

Strike 1.5

User Manual

Copyright © 2006 Schrödinger, LLC. All rights reserved. CombiGlide, Epik, Glide, Impact, Jaguar, Liaison, LigPrep, Maestro, Phase, Prime, QikProp, QikFit, QikSim, QSite, SiteMap, and Strike are trademarks of Schrödinger, LLC.

Schrödinger and MacroModel are registered trademarks of Schrödinger, LLC.

Python is a copyrighted work of the Python Software Foundation. All rights reserved.

To the maximum extent permitted by applicable law, this publication is provided “as is” without warranty of any kind. This publication may contain trademarks of other companies.

Please note that any third party programs (“Third Party Programs”) or third party Web sites (“Linked Sites”) referred to in this document may be subject to third party license agreements and fees. Schrödinger, LLC and its affiliates have no responsibility or liability, directly or indirectly, for the Third Party Programs or for the Linked Sites or for any damage or loss alleged to be caused by or in connection with use of or reliance thereon. Any warranties that we make regarding our own products and services do not apply to the Third Party Programs or Linked Sites, or to the interaction between, or interoperability of, our products and services and the Third Party Programs. Referrals and links to Third Party Programs and Linked Sites do not constitute an endorsement of such Third Party Programs or Linked Sites.

Revision A, April 2006

Contents

Document Conventions	vii
Chapter 1: Introduction to Strike	1
1.1 Strike Overview	1
1.2 Citing Strike in Publications	1
Chapter 2: Introduction to Maestro	3
2.1 General Interface Behavior	3
2.2 Starting Maestro	3
2.3 The Maestro Main Window	4
2.3.1 The Menu Bar	6
2.3.2 The Toolbar	7
2.3.3 Mouse Functions in the Workspace	10
2.3.4 Shortcut Key Combinations	11
2.4 Maestro Projects	11
2.4.1 The Project Table Toolbar	13
2.4.2 The Project Table Menus	14
2.4.3 Selecting Entries	15
2.4.4 Including Entries in the Workspace	16
2.4.5 Mouse Functions in the Project Table	16
2.4.6 Project Table Shortcut Keys	17
2.5 Building a Structure	18
2.5.1 Placing and Connecting Fragments	18
2.5.2 Adjusting Properties	20
2.5.3 The Build Panel Toolbar	20
2.6 Selecting Atoms	21
2.6.1 Toolbar Buttons	21
2.6.2 Picking Tools	22
2.6.3 The Atom Selection Dialog Box	23
2.7 Scripting in Maestro	23
2.7.1 Python Scripts	23

2.7.2 Command Scripts	24
2.7.3 Macros	25
2.8 Specifying a Maestro Working Directory	25
2.9 Undoing an Operation	26
2.10 Running and Monitoring Jobs	26
2.11 Getting Help	28
2.12 Ending a Maestro Session	28
Chapter 3: Strike Tutorial.....	29
3.1 Creating a Working Directory	29
3.2 Generating and Testing a QSPR Model for Aqueous Solubility	30
3.2.1 Starting Maestro	30
3.2.2 Importing Data	31
3.2.3 Preparing Test and Training Sets	33
3.2.4 Building a Partial Least Squares Model	34
3.2.5 Examining PLS Model-Building Results.....	36
3.2.6 Applying the Model to the Test Set	41
3.2.7 Calculating Univariate and Bivariate Statistics.....	45
3.2.8 Model-Building Using Principal Component Analysis	48
3.2.9 Model-Building Using Multiple Linear Regression	49
3.3 Calculating Atom-Pair Similarities	51
3.3.1 Changing Maestro Directories	51
3.3.2 Importing Active and Decoy Ligands	51
3.3.3 Opening the Calculate Similarity Panel.....	52
3.3.4 Seeding the Database and Designating Probes	53
3.3.5 Applying Atom-Pair Similarity.....	55
3.4 Calculating Descriptor Similarities from Molecular Properties	58
3.5 Estimating Activity by Creating a QSAR Model.....	60
3.5.1 Thymidylate Synthase Activity of Folate-Based Inhibitors	60
3.5.2 Changing Maestro Directories	61
3.5.3 Preparing the Data.....	61

3.5.4 Model Generation	62
3.5.5 Applying the Model to the Test Set	64
Chapter 4: Running Strike from Maestro	69
4.1 The Build QSAR Model Panel.....	69
4.1.1 Using the Build QSAR Model Panel.....	69
4.1.2 Build QSAR Model Panel Features.....	71
4.2 The Predict Based on QSAR Model Panel	73
4.2.1 Using the Predict Panel.....	73
4.2.2 Predict Panel Features.....	73
4.3 The Calculate Similarity Panel	74
4.3.1 Using the Calculate Similarity Panel	74
4.3.2 Calculate Similarity Panel Features	74
Chapter 5: Running Strike from the Command Line.....	77
5.1 Usage Summary	77
5.2 Input File Examples	77
5.3 Input File Keywords, Values, Descriptions	78
5.3.1 Mode Selection	78
5.3.2 File Specification Commands	78
5.3.3 Alternative Naming Convention Commands	79
5.3.4 Commands for Reading/Writing .CSV Files.....	79
5.3.5 Commands for Build QSAR Model (train) Jobs.....	80
5.3.6 Commands for Atom-Pair Similarity (apsimil) Jobs.....	81
5.3.7 Commands for Factor Reduction Jobs.....	82
5.3.8 Other Commands.....	82
5.3.9 Keyword Requirements for Various Job Types.....	82
Chapter 6: Statistical Definitions and Methods.....	85
6.1 Univariate Statistics	85
6.1.1 Symbols	85
6.1.2 Mean, Median, and Mode	85

6.1.3 Variance and Deviation	86
6.1.4 Skewness and Kurtosis	87
6.2 Bivariate Statistics: Covariance and Correlation	88
6.3 Model-Building Methods	90
6.3.1 Independent and Dependent Variables	91
6.3.2 Partial Least Squares	91
6.3.3 Principal Component Analysis	91
6.3.4 Multiple Linear Regression	92
6.4 Model Analysis and Validation	92
6.5 Outlier Detection	93
6.6 Similarity Statistics	94
6.6.1 Atom-Pair Similarity	94
6.6.2 Similarity Measures in Descriptor Space	94
Chapter 7: Getting Help	97
Index	99

Document Conventions

In addition to the use of italics for names of documents, the font conventions that are used in this document are summarized in the table below.

Table 3.1.

Font	Example	Use
Sans serif	Project Table	Names of GUI features, such as panels, menus, menu items, buttons, and labels
Monospace	\$SCHRODINGER/maestro	File names, directory names, commands, environment variables, and screen output
Italic	<i>filename</i>	Text that the user must replace with a value
Sans serif uppercase	CTRL+H	Keyboard keys

In descriptions of command syntax, the following UNIX conventions are used: braces { } enclose a choice of required items, square brackets [] enclose optional items, and the pipe symbol | separates items in a list from which one item must be chosen. Lines of command syntax that wrap should be interpreted as a single command.

In this document, to *type* text means to type the required text in the specified location, and to *enter* text means to type the required text, then press the ENTER key.

References to literature sources are given in square brackets, like this: [10].

Introduction to Strike

1.1 Strike Overview

Strike™ is a chemically-aware statistical package which is integrated with Maestro™ to provide a flexible and intuitive interface. Employing molecular data generated by Schrödinger software such as QikProp™, Glide™, Liaison™, or MacroModel®, or from other sources such as experimental data or third-party software, Strike can be used to do the following:

- Generate basic univariate and bivariate statistics such as mean, median, mode, covariance, and correlations
- Generate structure-activity relationship hypotheses using rigorous statistical methods
- Run validation tools to assess the validity and predictive power of generated QSAR/QSPR models
- Employ such models as filters and predictive tools
- Perform similarity analysis in molecular property or 2-dimensional structural space.

This document provides an introduction to Maestro, a set of tutorial exercises using the capabilities of Strike, a description of the Strike GUI in Maestro, a command line reference chapter, and definitions of some statistics terms and methods.

1.2 Citing Strike in Publications

The use of this program should be acknowledged in publications as:

Strike, version 1.5, Schrödinger, LLC, New York, NY, 2005.

Introduction to Maestro

Maestro is the graphical user interface for all of Schrödinger's products: CombiGlide™, Epik™, Glide™, Impact™, Jaguar™, Liaison™, LigPrep™, MacroModel®, Phase™, Prime™, QikProp™, QSite™, and Strike™. It contains tools for building, displaying, and manipulating chemical structures; for organizing, loading, and storing these structures and associated data; and for setting up, monitoring, and visualizing the results of calculations on these structures. This chapter provides a brief introduction to Maestro and some of its capabilities. For more information on any of the topics in this chapter, see the [Maestro User Manual](#).

2.1 General Interface Behavior

Most Maestro panels are amodal: more than one panel can be open at a time, and a panel need not be closed for an action to be carried out. Each Maestro panel has a Close button so you can hide the panel from view.

Maestro supports the mouse functions common to many graphical user interfaces. The left button is used for choosing menu items, clicking buttons, and selecting objects by clicking or dragging. This button is also used for resizing and moving panels. The right button displays a shortcut menu. Other common mouse functions are supported, such as using the mouse in combination with the SHIFT or CTRL keys to select a range of items and select or deselect a single item without affecting other items.

In addition, the mouse buttons are used for special functions described later in this chapter. These functions assume that you have a three-button mouse. If you have a two-button mouse, ensure that it is configured for three-button mouse simulation (the middle mouse button is simulated by pressing or holding down both buttons simultaneously).

2.2 Starting Maestro

Before starting Maestro, you must first set the SCHRODINGER environment variable to point to the installation directory. To set this variable, enter the following command at a shell prompt:

```
cshtcsh: setenv SCHRODINGER installation-directory
bashksh: export SCHRODINGER=installation-directory
```

You might also need to set the `DISPLAY` environment variable, if it is not set automatically when you log in. To determine if you need to set this variable, enter the command:

```
echo $DISPLAY
```

If the response is a blank line, set the variable by entering the following command:

```
csh/tcsh:      setenv DISPLAY display-machine-name:0.0
```

```
bash/ksh:      export DISPLAY=display-machine-name:0.0
```

After you set the `SCHRODINGER` and `DISPLAY` environment variables, you can start Maestro using the command:

```
$SCHRODINGER/maestro options
```

If you add the `$SCHRODINGER` directory to your path, you only need to enter the command `maestro`. Options for this command are given in [Section 2.1](#) of the *Maestro User Manual*.

The directory from which you started Maestro is Maestro's current working directory, and all data files are written to and read from this directory unless otherwise specified (see [Section 2.8 on page 25](#)). You can change directories by entering the following command in the command input area (see [page 6](#)) of the main window:

```
cd directory-name
```

where *directory-name* is either a full path or a relative path.

2.3 The Maestro Main Window

The Maestro main window is shown in [Figure 2.1 on page 5](#). The main window components are listed below.

The following components are always visible:

- **Title bar**—displays the Maestro version, the project name (if there is one) and the current working directory.
- **Auto-Help**—automatically displays context-sensitive help.
- **Menu bar**—provides access to panels.
- **Workspace**—displays molecular structures and other 3D graphical objects.

The following components can be displayed or hidden by choosing the component from the Display menu. Your choice of which main window components are displayed is persistent between Maestro sessions.

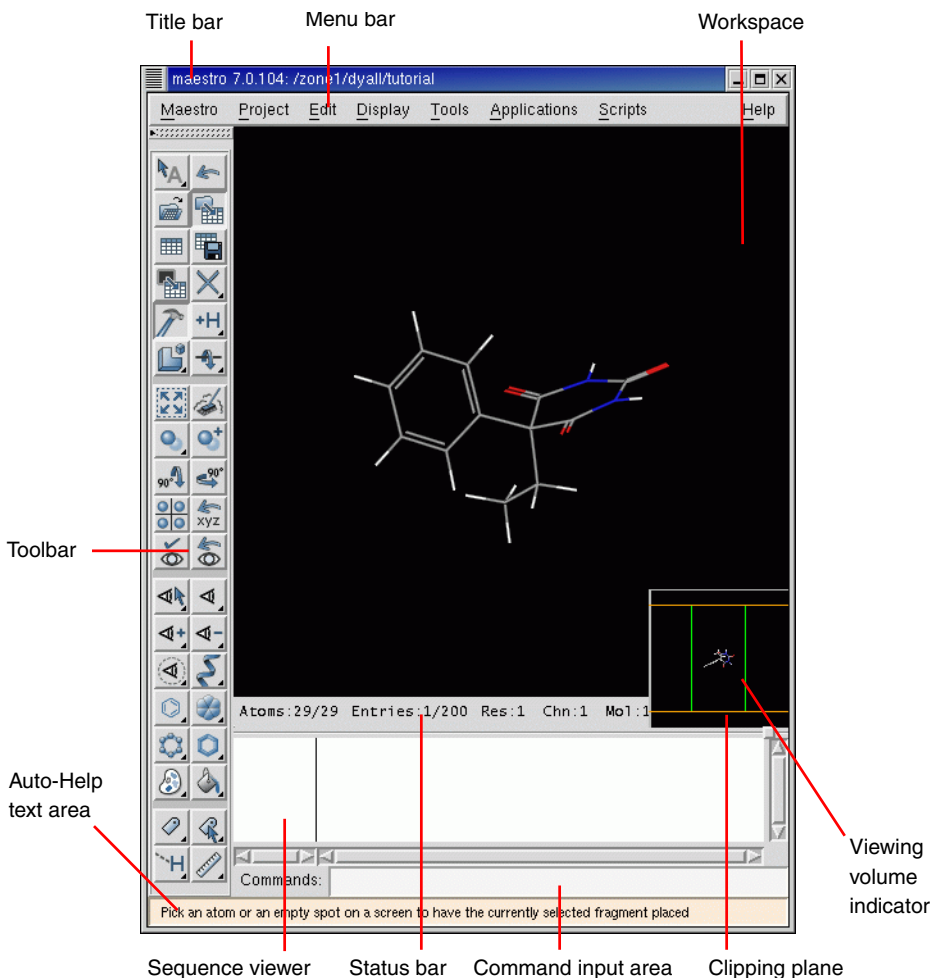


Figure 2.1. The Maestro main window.

- **Toolbar**—contains buttons for many common tasks and provides tools for displaying and manipulating structures, as well as organizing the Workspace.
- **Status bar**—displays information about a particular atom, or about structures in the Workspace, depending on where the pointer pauses (see [Section 2.5](#) of the *Maestro User Manual* for details):
 - **Atom**—displays the chain, residue number, element, PDB atom name, formal charge, and title or entry name (this last field is set by choosing Preferences from the Maestro menu and selecting the Feedback folder).

- **Workspace**—displays the number of atoms, entries, residues, chains, and molecules in the Workspace.
- **Clipping planes window**—displays a small, top view of the Workspace and shows the clipping planes and viewing volume indicators.
- **Sequence viewer**—shows the sequences for proteins displayed in the Workspace. See [Section 2.6](#) of the *Maestro User Manual* for details.
- **Command input area**—provides a place to enter Maestro commands.

When a distinction between components in the main window and those in other panels is needed, the term *main* is applied to the main window components (e.g., main toolbar).

You can expand the Workspace to occupy the full screen, by pressing CTRL+=. All other components and panels are hidden. To return to the previous display, press CTRL+= again.

2.3.1 The Menu Bar

The menus on the main menu bar provide access to panels, allow you to execute commands, and control the appearance of the Workspace. The main menus are as follows:

- **Maestro**—save or print images in the Workspace, execute system commands, save or load a panel layout, set preferences, set up Maestro command aliases, and quit Maestro.
- **Project**—open and close projects, import and export structures, make a snapshot, and annotate a project. These actions can also be performed from the Project Table panel. For more information, see [Section 2.4 on page 11](#).
- **Edit**—undo actions, build and modify structures, define command scripts and macros, and find atoms in the Workspace.
- **Display**—control the display of the contents of the Workspace, arrange panels, and display or hide main window components.
- **Tools**—group atoms; measure, align, and superimpose structures; and view and visualize data.
- **Applications**—set up, submit, and monitor jobs for Schrödinger’s computational programs. Some products have a submenu from which you can choose the task to be performed.
- **Scripts**—manage and install Python scripts that come with the distribution and scripts that you create yourself. (See [Chapter 13](#) of the *Maestro User Manual* for details.)
- **Help**—open the Help panel, the PDF documentation index, or information panels; run a demonstration; and display or hide Balloon Help (tooltips).

2.3.2 The Toolbar

The main toolbar contains three kinds of buttons for performing common tasks:



Action—Perform a simple task, like clearing the Workspace.



Display—Open or close a panel or open a dialog box, such as the Project Table panel.



Menu—Display a *button menu*. These buttons have a triangle in the lower right corner.

There are four types of items on button menus, and all four types can be on the same menu (see Figure 2.2):

- **Action**—Perform an action immediately.
- **Display**—Open a panel or dialog box.
- **Object types for selection**—Choose Atoms, Bonds, Residues, Chains, Molecules, or Entries, then click on an atom in the Workspace to perform the action on all the atoms in that structural unit.

The object type is marked on the menu with a red diamond and the button is indented to indicate the action to be performed.

- **Other setting**—Set a state, choose an attribute, or choose a parameter and click on atoms in the Workspace to display or change that parameter.

The toolbar buttons are described below. Some descriptions refer to features not described in this chapter. See the *Maestro User Manual* for a fuller description of these features.

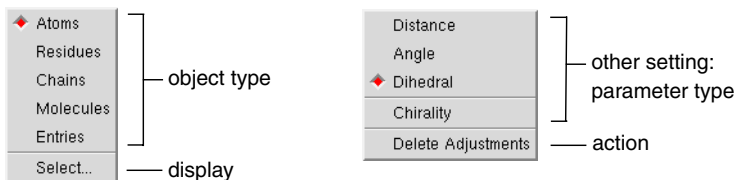


Figure 2.2. The Workspace selection *button menu* and the Adjust distances, angles or dihedrals *button menu*.

Workspace selection

- Choose an object type for selecting
- Open the Atom Selection dialog box

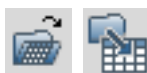


Undo/Redo

Undo or redo the last action. Performs the same function as the Undo item on the Edit menu, and changes to an arrow pointing in the opposite direction when an Undo has been performed, indicating that its next action is Redo.

Open a project

Open the Open Project dialog box.



Import structures

Open the Import panel.

Open/Close Project Table

Open the Project Table panel or close it if it is open.



Save as

Open the Save Project As dialog box, to save the project with a new name.

Create entry from Workspace

Open a dialog box in which you can create an entry in the current project using the contents of the Workspace.

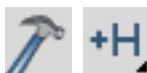


Delete

- Choose an object type for deletion
- Delete hydrogens and waters
- Open the Atom Selection dialog box
- Delete other items associated with the structures in the Workspace
- Click to select atoms to delete
- Double-click to delete all atoms

Open/Close Build panel

Open the Build panel or close it if it is open.



Add hydrogens

- Choose an object type for applying a hydrogen treatment
- Open the Atom Selection dialog box
- Click to select atoms to treat
- Double-click to apply to all atoms

Local transformation

- Choose an object type for transforming
- Click to select atoms to transform
- Open the Advanced Transformations panel



Adjust distances, angles or dihedrals

- Choose a parameter for adjusting
- Delete adjustments

Fit to screen

Scale the displayed structure to fit into the Workspace and reset the center of rotation.



Clear Workspace

Clear all atoms from the Workspace.

Set fog display state

Choose a fog state. Automatic means fog is on when there are more than 40 atoms in the Workspace, otherwise it is off.



Enhance depth cues

Optimize fogging and other depth cues based on what is in the Workspace.

Rotate around X axis by 90 degrees

Rotate the Workspace contents around the X axis by 90 degrees.



Rotate around Y axis by 90 degrees

Rotate the Workspace contents around the Y axis by 90 degrees.

Tile entries

Arrange entries in a rectangular grid in the Workspace.

**Save view**

Save the current view of the Workspace: orientation, location, and zoom.

**Display only selected atoms**

- Choose an object type for displaying
- Click to select atoms to display
- Double-click to display all atoms

**Also display**

- Choose a predefined atom category
- Open the Atom Selection dialog box

**Display residues within N angstroms of currently displayed atoms**

- Choose a radius
- Open a dialog box to set a value

**Draw bonds in wire**

- Choose an object type for drawing bonds in wire representation
- Open the Atom Selection dialog box
- Click to select atoms for representation
- Double-click to apply to all atoms

**Draw atoms in Ball & Stick**

- Choose an object type for drawing bonds in Ball & Stick representation
- Open the Atom Selection dialog box
- Click to select atoms for representation
- Double-click to apply to all atoms

**Color all atoms by scheme**

Choose a predefined color scheme.

**Label atoms**

- Choose a predefined label type
- Delete labels

**Reset Workspace**

Reset the rotation, translation, and zoom of the Workspace to the default state.

**Restore view**

Restore the last saved view of the Workspace: orientation, location, and zoom.

**Display only**

- Choose a predefined atom category
- Open the Atom Selection dialog box

**Undisplay**

- Choose a predefined atom category
- Open the Atom Selection dialog box

**Show, hide, or color ribbons**

- Choose to show or hide ribbons
- Choose a color scheme for coloring ribbons

**Draw atoms in CPK**

- Choose an object type for drawing bonds in CPK representation
- Open the Atom Selection dialog box
- Click to select atoms for representation
- Double-click to apply to all atoms

**Draw bonds in tube**

- Choose an object type for drawing bonds in tube representation
- Open the Atom Selection dialog box
- Click to select atoms for representation
- Double-click to apply to all atoms

**Color residue by constant color**

- Choose a color for applying to residues
- Click to select residues to color
- Double-click to color all atoms

**Label picked atoms**

- Choose an object type for labeling atoms
- Open the Atom Selection dialog box
- Open the Atom Labels panel at the Composition folder
- Delete labels
- Click to select atoms to label
- Double-click to label all atoms



Display H-bonds

- Choose bond type:
intra—displays H-bonds within the selected molecule
- inter—displays H-bonds between the selected molecule and all other atoms.
- Delete H-bonds
- Click to select molecule



Measure distances, angles or dihedrals

- Choose a parameter for displaying measurements
- Delete measurements
- Click to select atoms for measurement

2.3.3 Mouse Functions in the Workspace

The left mouse button is used for selecting objects. You can either click on a single atom or bond, or you can drag to select multiple objects. The right mouse button opens shortcut menus, which are described in [Section 2.7](#) of the *Maestro User Manual*.

The middle and right mouse buttons can be used on their own and in combination with the SHIFT and CTRL keys to perform common operations, such as rotating, translating, centering, adjusting, and zooming.

Table 2.1. Mapping of Workspace operations to mouse actions.

Mouse Button	Keyboard	Motion	Action
Left		click, drag	Select
Left	SHIFT	click, drag	Toggle the selection
Middle		drag	Rotate about X and Y axes Adjust bond, angle, or dihedral
Middle	SHIFT	drag vertically	Rotate about X axis
Middle	SHIFT	drag horizontally	Rotate about Y axis
Middle	CTRL	drag horizontally	Rotate about Z axis
Middle	SHIFT + CTRL	drag horizontally	Zoom
Right		click	Spot-center on selection
Right		click and hold	Display shortcut menu
Right		drag	Translate in the X-Y plane
Right	SHIFT	drag vertically	Translate along the X axis
Right	SHIFT	drag horizontally	Translate along the Y axis
Right	CTRL	drag horizontally	Translate along the Z axis
Middle & Right		drag horizontally	Zoom

2.3.4 Shortcut Key Combinations

Some frequently used operations have been assigned shortcut key combinations. The shortcuts available in the main window are described in [Table 2.2](#).

Table 2.2. Shortcut keys in the Maestro main window.

Keys	Action	Equivalent Menu Choices
CTRL+B	Open Build panel	Edit > Build
CTRL+C	Create entry	Project > Create Entry From Workspace
CTRL+E	Open Command Script Editor panel	Edit > Command Script Editor
CTRL+F	Open Find Atoms panel	Edit > Find
CTRL+H	Open Help panel	Help > Help
CTRL+I	Open Import panel	Project > Import Structures
CTRL+M	Open Measurements panel	Tools > Measurements
CTRL+N	Create new project	Project > New
CTRL+O	Open project	Project > Open
CTRL+P	Print	Maestro > Print
CTRL+Q	Quit	Maestro > Quit
CTRL+S	Open Sets panel	Tools > Sets
CTRL+T	Open Project Table panel	Project > Show Table
CTRL+W	Close project	Project > Close
CTRL+Z	Undo/Redo last command	Edit > Undo/Redo
CTRL+=	Enter and exit full screen mode (Workspace occupies full screen)	None

2.4 Maestro Projects

All the work you do in Maestro is done within a *project*. A project consists of a set of *entries*, each of which contains one or more chemical structures and their associated data. In any Maestro session, there can be only one Maestro project open. If you do not specify a project when you start Maestro, a *scratch* project is created. You can work in a scratch project without saving it, but you must save it in order to use it in future sessions. When you save or close a project, all the view transformations (rotation, translation, and zoom) are saved with it. When you close a project, a new scratch project is automatically created.

Likewise, if there is no entry displayed in the Workspace, Maestro creates a *scratch* entry. Structures that you build in the Workspace constitute a scratch entry until you save the structures as project entries. The scratch entry is not saved with the project unless you explicitly add it to the project. However, you can use a scratch entry as input for some calculations.

To add a scratch entry to a project, do one of the following:

- Click the Create entry from Workspace button:



- Choose Create Entry from Workspace from the Project menu.
- Press CTRL+C.

In the dialog box, enter a name and a title for the entry. The entry name is used internally to identify the entry and can be modified by Maestro. The title can be set or changed by the user, but is not otherwise modified by Maestro.

Once an entry has been incorporated into the project, its structures and their data are represented by a row in the Project Table. Each row contains the row number, an icon indicating whether the entry is displayed in the Workspace (the In column), the entry title, a button to open the Surfaces panel if the entry has surfaces, the entry name, and any entry properties. The row number is not a property of the entry.

Entries can be collected into groups, and the members of the group can be displayed or hidden. Most additions of multiple entries to the Project Table are done as entry groups.

You can use entries as input for all of the computational programs—Glide, Impact, Jaguar, Liaison, LigPrep, MacroModel, Phase, Prime, QikProp, QSite, and Strike. You can select entries as input for the ePlayer, which displays the selected structures in sequence. You can also duplicate, combine, rename, and sort entries; create properties; import structures as entries; and export structures and properties from entries in various formats.

To open the Project Table panel, do one of the following:

- Click the Open/Close Project Table button on the toolbar



- Choose Show Table from the Project menu
- Press CTRL+T.

The Project Table panel contains a menu bar, a toolbar, and the table itself.

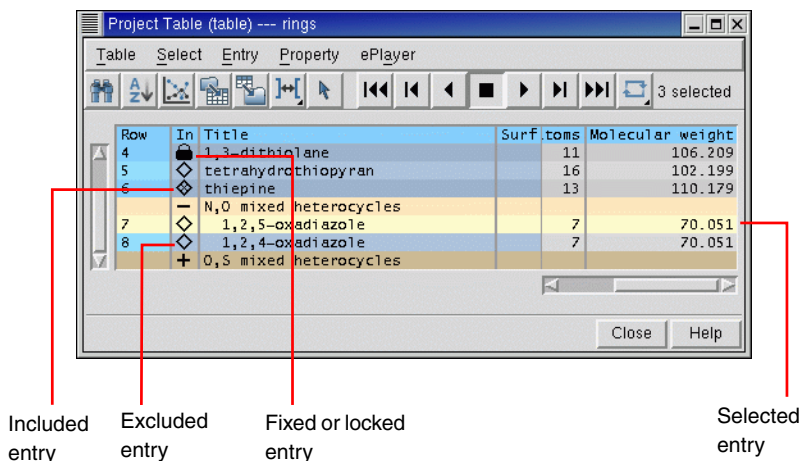


Figure 2.3. The Project Table panel.

2.4.1 The Project Table Toolbar

The Project Table toolbar contains two groups of buttons and a status display. The first set of buttons opens various panels that allow you to perform functions on the entries in the Project Table. The second set of buttons controls the ePlayer, which “plays through” the selected structures: each structure is displayed in the Workspace in sequence, at a given time interval. See [Section 2.3.2 on page 7](#) for a description of the types of toolbar buttons. The buttons are described below.



Find

Open the Find panel for locating alphanumeric text in any column of the Project Table, except for the row number.



Sort

Open the Sort panel for sorting entries by up to three properties.



Plot

Open the Plot panel for plotting entry properties.



Import Structure

Open the Import panel for importing structures into the project.



Export Structure

Open the Export panel for exporting structures to a file.



Columns

Choose an option for adjusting the column widths.



Select only

Open the Entry Selection dialog box for selecting entries based on criteria for entry properties.



Go to start

Display the first selected structure.



Previous

Display the previous structure in the list of selected structures.



Play backward

Display the selected structures in sequence, moving toward the first.



Stop

Stop the ePlayer.



Play forward

Display the selected structures in sequence, moving toward the last.



Next

Display the next structure in the list of selected structures.



Go to end

Display the last selected structure.



Loop

Choose an option for repeating the display of the structures. **Single Direction** displays structures in a single direction, then repeats. **Oscillate** reverses direction each time the beginning or end of the list is reached.

The status display, to the right of the toolbar buttons, shows the number of selected entries. When you pause the cursor over the status display, the Balloon Help shows the total number of entries, the number shown in the table, the number selected, and the number included in the Workspace.

2.4.2 The Project Table Menus

- **Table**—find text, sort entries, plot properties, import and export structures, and configure the Project Table.
- **Select**—select all entries, none, invert your selection, or select classes of entries using the Entry Selection dialog box and the Filter panel.
- **Entry**—include or exclude entries from the Workspace, display or hide entries in the Project Table, and perform various operations on the selected entries.


- **Property**—display and manipulate entry properties in the Project Table.
- **ePlayer**—view entries in succession, stop, reverse, and set the ePlayer options.

2.4.3 Selecting Entries

Many operations in Maestro are performed on the entries selected in the Project Table. The Project Table functions much like any other table: select rows by clicking, shift-clicking, and control-clicking. However, because clicking in an editable cell of a selected row enters edit mode, you should click in the Row column to select entries. See [Section 2.4.5 on page 16](#) for more information on mouse actions in the Project Table. There are shortcuts for selecting classes of entries on the **Select** menu.

In addition to selecting entries manually, you can select entries that meet a combination of conditions on their properties. Such combinations of conditions are called *filters*. Filters are Entry Selection Language (ESL) expressions and are evaluated at the time they are applied. For example, if you want to set up a Glide job that uses ligands with a low molecular weight (say, less than 300) and that has certain QikProp properties, you can set up a filter and use it to select entries for the job. If you save the filter, you can use it again on a different set of ligands that meet the same selection criteria.

To create a filter:

1. Do one of the following:
 - Choose **Only**, **Add**, or **Deselect** from the **Select** menu.
 - Click the **Entry selection** button on the toolbar.
- 
2. In the **Properties** folder, select a property from the property list, then select a condition.
 3. Combine this selection with the current filter by clicking **Add**, **Subtract**, or **Intersect**. These buttons perform the Boolean operations **OR**, **AND NOT**, and **AND** on the corresponding ESL expressions.
 4. To save the filter for future use click **Create Filter**, enter a name, and click **OK**.
 5. Click **OK** to apply the filter immediately.

2.4.4 Including Entries in the Workspace

In addition to selecting entries, you can also use the Project Table to control which entries are displayed in the Workspace. An entry that is displayed in the Workspace is *included* in the Workspace; likewise, an entry that is not displayed is *excluded*. Included entries are marked by an X in the diamond in the In column; excluded entries are marked by an empty diamond. Entry inclusion is completely independent of entry selection.

To include or exclude entries, click, shift-click, or control-click in the In column of the entries, or select entries and choose Include or Exclude from the Entry menu. Inclusion with the mouse works just like selection: when you include an entry by clicking, all other entries are excluded.

It is sometimes useful to keep one entry in the Workspace and include others one by one: for example, a receptor and a set of ligands. You can fix the receptor in the Workspace by selecting it in the Project Table and choosing Fix from the Entry menu or by pressing CTRL+F. A padlock icon replaces the diamond in the In column to denote a *fixed* entry. To remove a fixed entry from the Workspace, you must exclude it explicitly (CTRL+X). It is not affected by the inclusion or exclusion of other entries. Fixing an entry affects only its inclusion; you can still rotate, translate, or modify the structure.

2.4.5 Mouse Functions in the Project Table

The Project Table supports the standard use of shift-click and control-click to select objects. This behavior applies to the selection of entries and the inclusion of entries in the Workspace. You can also drag to resize rows and columns and to move rows.

You can drag a set of non-contiguous entries to reposition them in the Project Table. When you release the mouse button, the entries are placed after the first unselected entry that precedes the entry on which the cursor is resting. For example, if you select entries 2, 4, and 6, and release the mouse button on entry 3, these three entries are placed after entry 1, because entry 1 is the first unselected entry that precedes entry 3. To move entries to the top of the table, drag them above the top of the table; to move entries to the end of the table, drag them below the end of the table.

A summary of mouse functions in the Project Table is provided in [Table 2.3](#).

Table 2.3. Mouse operations in the Project Table.

Task	Mouse Operation
Change a Boolean property value	Click repeatedly in a cell to cycle through the possible values (On, Off, Clear)
Display the Entry menu for an entry	Right-click anywhere in the entry. If the entry is not selected, it becomes the selected entry. If the entry is selected, the action is applied to all selected entries.
Display a version of the Property menu for a property	Right-click in the column header
Edit the text or the value in a table cell	Click in the cell and edit the text or value
Include an entry in the Workspace, exclude all others	Click the In column of the entry
Move selected entries	Drag the entries
Paste text into a table cell	Middle-click
Resize rows or columns	Drag the boundary with the middle mouse button
Select an entry, deselect all others	For an unselected entry, click anywhere in the row except the In column; for a selected entry, click the row number.
Select or include multiple entries	Click the first entry then shift-click the last entry
Toggle the selection or inclusion state	Control-click the entry or the In column

2.4.6 Project Table Shortcut Keys

Some frequently used project operations have been assigned shortcut key combinations. The shortcuts, their functions, and their menu equivalents are listed in [Table 2.4](#).

Table 2.4. Shortcut keys in the Project Table.

Keys	Action	Equivalent Menu Choices
CTRL+A	Select all entries	Select > All
CTRL+F	Fix entry in Workspace	Entry > Fix
CTRL+I	Open Import panel	Table > Import Structures
CTRL+N	Include only selected entries	Entry > Include Only
CTRL+U	Deselect all entries	Select > None
CTRL+X	Exclude selected entries	Entry > Exclude
CTRL+Z	Undo/Redo last command	Edit > Undo/Redo in main window

2.5 Building a Structure

After you start Maestro, the first task is usually to create or import a structure. You can open existing Maestro projects or import structures from other sources to obtain a structure, or you can build your own. To open the Build panel, do one of the following:

- Click the Open/Close Build panel button in the toolbar:



- Choose Build from the Edit menu.
- Press CTRL+B.

The Build panel allows you to create structures by drawing or placing atoms or fragments in the Workspace and connecting them into a larger structure, to adjust atom positions and bond orders, and to change atom properties. This panel contains a toolbar and three folders.

2.5.1 Placing and Connecting Fragments

The Build panel provides several tools for creating structures in the Workspace. You can place and connect fragments, or you can draw a structure freehand.

To place a fragment in the Workspace:

1. Select Place.
2. Choose a fragment library from the Fragments menu.
3. Click a fragment.
4. Click in the Workspace where you want the fragment to be placed.

To connect fragments in the Workspace, do one of the following:

- Place another fragment and connect them using the Connect & Fuse panel, which you open from the Edit menu on the main menu bar or with the Display Connect & Fuse panel on the Build toolbar.



- Replace one or more atoms in the existing fragment with another fragment by selecting a fragment and clicking in the Workspace on the main atom to be replaced.
- Grow another fragment by selecting Grow in the Build panel and clicking the fragment you want to add in the Fragments folder.

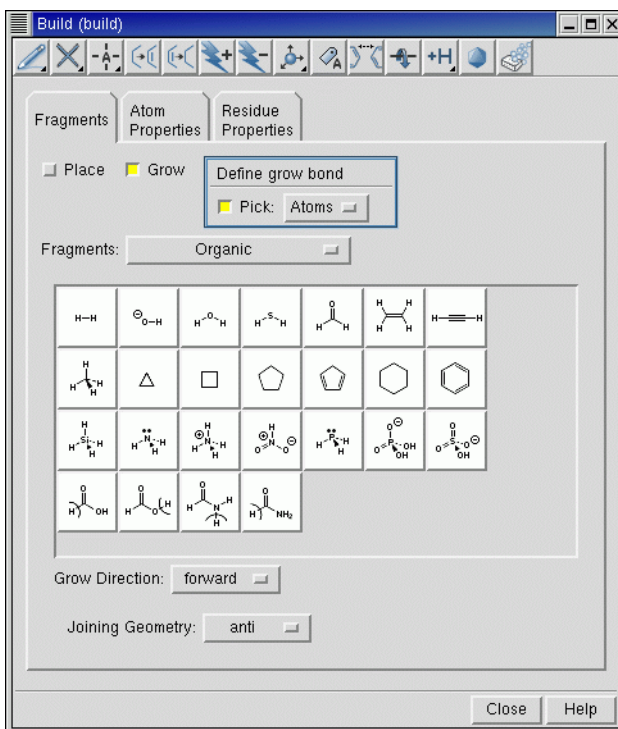


Figure 2.4. The Build panel.

Grow mode uses predefined rules to connect a fragment to the *grow bond*. The grow bond is marked by a green arrow. The new fragment replaces the atom at the head of the arrow on the grow bond and all atoms attached to it. To change the grow bond, choose Bonds from the Pick option menu in the Build panel and click on the desired grow bond in the Workspace. The arrow points to the atom nearest to where you clicked.

To draw a structure freehand:

1. Choose an element from the Draw button menu on the Build panel toolbar:



2. Click in the Workspace to place an atom of that element.
3. Click again to place another atom and connect it to the previous atom.
4. Continue this process until you have drawn the structure.
5. Click the active atom again to finish drawing.

2.5.2 Adjusting Properties

In the Atom Properties folder, you can change the properties of the atoms in the Workspace. For each item on the Property option menu—Element, Atom Type (MacroModel), Partial Charge, PDB Atom Name, Grow Name, and Atom Name—there is a set of tools you can use to change the atom properties. For example, the Element tools consist of a periodic table from which you can choose an element and select an atom to change it to an atom of the selected element.

Similarly, the Residue Properties folder provides tools for changing the properties of residues: the Residue Number, the Residue Name, and the Chain Name.

To adjust bond lengths, bond angles, dihedral angles, and chiralities during or after building a structure, use the Adjust distances, angles or dihedrals button on the main toolbar:



You can also open the Adjust panel from this button menu, from the Display Adjust panel button on the Build panel toolbar (which has the same appearance as the above button) or from the Edit menu in the main window.

2.5.3 The Build Panel Toolbar

The toolbar of the Build panel provides quick access to tools for drawing and modifying structures and labeling atoms. See [Section 2.3.2 on page 7](#) for a description of the types of toolbar buttons. The toolbar buttons and their use are described below.



Free-hand drawing

Choose an element for drawing structures freehand in the Workspace (default C). Each click in the Workspace places an atom and connects it to the previous atom.



Delete

Choose an object for deleting. Same as the [Delete](#) button on the main toolbar, see [page 8](#).



Set element

Choose an element for changing atoms in the Workspace (default C). Click an atom to change it to the selected element.



Increment bond order

Select a bond to increase its bond order by one, to a maximum of 3.



Decrement bond order

Select a bond to decrease its bond order by one, to a minimum of 0.

**Increment formal charge**

Select an atom to increase its formal charge by one.

**Decrement formal charge**

Select an atom to decrease its formal charge by one.

**Move**

Choose a direction for moving atoms, then click the atom to be moved. Moves in the XY plane are made by clicking the new location. Moves in the Z direction are made in 0.5 Å increments.

**Label**

Apply heteroatom labels as you build a structure. The label consists of the element name and formal charge, and is applied to atoms other than C and H.

**Display Connect & Fuse panel**

Open the Connect & Fuse panel so you can connect structures (create bonds between structures) or fuse structures (replace atoms of one structure with those of another).

**Display Adjust panel**

Open the Adjust panel so you can change bond lengths, bond angles, dihedral angles, or atom chiralities.

**Add hydrogens**

Choose an atom type for applying the current hydrogen treatment. Same as the [Add hydrogens](#) button on the main toolbar, see [page 8](#).

**Geometry Symmetrizer**

Open the Geometry Symmetrizer panel for symmetrizing the geometry of the structure in the Workspace.

**Geometry Cleanup**

Clean up the geometry of the structure in the Workspace.

2.6 Selecting Atoms

Maestro has a powerful set of tools for selecting atoms in a structure: toolbar buttons, picking tools in panels, and the Atom Selection dialog box. These tools allow you to select atoms in two ways:

- Select atoms first and apply an action to them
- Choose an action first and then select atoms for that action

2.6.1 Toolbar Buttons

The small triangle in the lower right corner of a toolbar button indicates that the button contains a menu. Many of these buttons allow you to choose an object type for selecting: choose Atoms, Bonds, Residues, Chains, Molecules, or Entries, then click on an atom in the Workspace to perform the action on all the atoms in that structural unit.

For example, to select atoms with the Workspace selection toolbar button:

1. Choose Residues from the Workspace selection button menu:



The button changes to:



2. Click on an atom in a residue in the Workspace to select all the atoms in that residue.

2.6.2 Picking Tools

The picking tools are embedded in each panel in which you need to select atoms to apply an operation. The picking tools in a panel can include one or more of the following:

- Pick option menu—Allows you to choose an object type. Depending on the operation to be performed, you can choose Atoms, Bonds, Residues, Chains, Molecules, or Entries, then click on an atom in the Workspace to perform the action on all the atoms in that structural unit.

The Pick option menu varies from panel to panel, because not all object types are appropriate for a given operation. For example, some panels have only Atoms and Bonds in the Pick option menu.

- All button—Performs the action on all atoms in the Workspace.
- Selection button—Performs the action on any atoms already selected in the Workspace.
- Previous button—Performs the action on the most recent atom selection defined in the Atom Selection dialog box.
- Select button—Opens the Atom Selection dialog box.
- ASL text box—Allows you to type in an ASL expression for selecting atoms.

ASL stands for Atom Specification Language, and is described in detail in the [Maestro Command Reference Manual](#).

- Clear button—Clears the current selection



- Show markers option—Marks the selected atoms in the Workspace.

For example, to label atoms with the Label Atoms panel:

1. Choose Atom Labels from the Display menu.
2. In the Composition folder, select Element and Atom Number.
3. In the picking tools section at the top of the panel, you could do one of the following:
 - Click Selection to apply labels to the atoms already selected in the Workspace (from the previous example).
 - Choose Residues from the Pick option menu and click on an atom in a different residue to label all the atoms in that residue.

2.6.3 The Atom Selection Dialog Box

If you wish to select atoms based on more complex criteria, you can use the Atom Selection dialog box. To open this dialog box, choose Select from a button menu or click the Select button in a panel. See [Section 5.3](#) of the *Maestro User Manual* for detailed instructions on how to use the Atom Selection dialog box.

2.7 Scripting in Maestro

Although you can perform nearly all Maestro-supported operations through menus and panels, you can also perform operations using Maestro commands, or compilations of these commands, called *scripts*. Scripts can be used to automate lengthy procedures or repetitive tasks and can be created in several ways. These are summarized below.

2.7.1 Python Scripts

Python is a full-featured scripting language that has been embedded in Maestro to extend its scripting facilities. The Python capabilities within Maestro include access to Maestro functionality for dealing with chemical structures, projects, and Maestro files.

The two main Python commands used in Maestro are:

- `pythonrun`—executes a Python module. (You can also use the alias `pyrun`.) The syntax is:

```
pythonrun module.function
```
- `pythonimport`—rereads a Python file so that the next time you use the `pythonrun` command, it uses the updated version of the module. (You can also use the alias `pyimp`.)

From the Maestro Scripts menu you can install, manage, and run Python scripts. For more information on the Scripts menu, see [Section 13.1](#) of the *Maestro User Manual*.

For more information on using Python with Maestro, see *Scripting with Python*.

2.7.2 Command Scripts

All Maestro commands are logged and displayed in the Command Script Editor panel. This means you can create a command script by performing the operations with the GUI controls, copying the logged commands from the Command History list into the Script text area of the panel, then saving the list of copied commands as a script.

To run an existing command script:

1. Open the Command Script Editor panel from the Edit menu in the main window.
2. Click Open Local and navigate to the directory containing the desired script.
3. Select a script in the Files list and click Open.

The script is loaded into the Script window of the Command Script Editor panel.

4. Click Run Script.

Command scripts cannot be used for Prime operations.

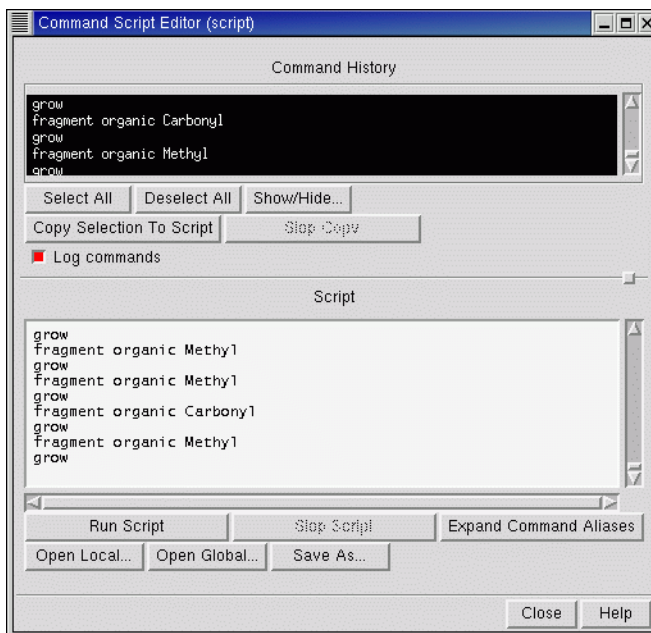


Figure 2.5. The Command Script Editor *panel*.

2.7.3 Macros

There are two kinds of macros you can create: named macros and macros assigned to function keys F1 through F12.

To create and run a named macro:

1. Open the Macros panel from the Edit menu in the main window.
2. Click New, enter a name for the macro, and click OK.
3. In the Definition text box, type the commands for the macro.
4. Click Update to update the macro definition.
5. To run the macro, enter the following in the command input area in the main window:

```
macrorun macro-name
```

If the command input area is not visible, choose Command Input Area from the Display menu.

To create and run a function key macro:

1. Open the Function Key Macros panel from the Edit menu in the main window.
2. From the Macro Key option, select a function key (F1 through F12) to which to assign the macro.
3. In the text box, type the commands for the macro.
4. Click Run to test the macro or click Save to save it.
5. To run the macro from the main window, press the assigned function key.

For more information on macros, see [Section 13.5](#) of the *Maestro User Manual*.

2.8 Specifying a Maestro Working Directory

When you use Maestro to launch Strike jobs, Maestro writes job output to the directory specified in the Directory folder of the Preferences panel. By default, this directory (the file I/O directory) is the directory from which you started Maestro.

To change the Maestro working directory:

1. Open the Preferences panel from the Maestro menu.
2. Click the Directory tab.
3. Select the directory you want to use for reading and writing files.

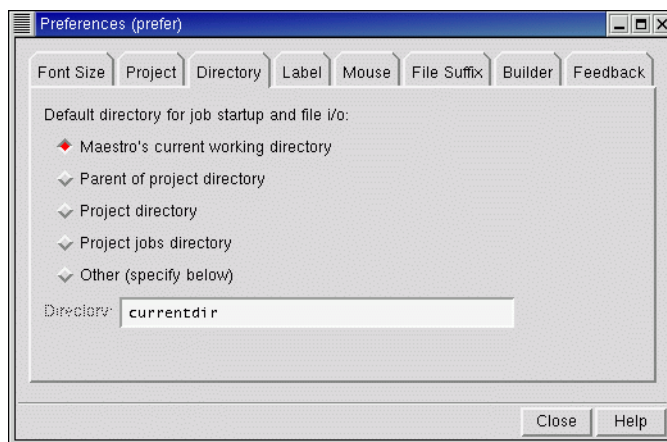


Figure 2.6. The Directory *folder of the* Preferences *panel*.

You can also set other preferences in the Preferences panel. See [Section 12.2](#) of the *Maestro User Manual* for details.

2.9 Undoing an Operation

To undo a single operation, click the Undo button in the toolbar, choose Undo from the Edit menu, or press CTRL+Z. The word Undo in the menu is followed by text that describes the operation to undo. Not all operations can be undone: for example, global rotations and translations are not undoable operations. For such operations you can use the Save view and Restore view buttons in the toolbar, which save and restore a molecular orientation.

2.10 Running and Monitoring Jobs

Maestro has panels for each product for preparing and submitting jobs. To use these panels, choose the appropriate product and task from the Applications menu and its submenus. Set the appropriate options in the panel, then click Start to open the Start dialog box and set options for running the job. For a complete description of the Start dialog box associated with your computational program, see your product's User Manual. When you have finished setting the options, click Start to launch the job and open the Monitor panel.

The Monitor panel is the control panel for monitoring the progress of jobs and for pausing, resuming, or killing jobs. All jobs that belong to you can be displayed in the Monitor panel, whether or not they were started from Maestro. Subjobs are indented under their parent in the job list. The text pane shows output information from the monitored job, such as the contents

of the log file. The Monitor panel opens automatically when you start a job. If it is not open, you can open it by choosing Monitor from the Applications menu in the Maestro main window.

While jobs are running, the Detach, Pause, Resume, Stop, Kill, and Update buttons are active. When there are no jobs currently running, only the Monitor and Delete buttons are active. These buttons act on the selected job. By default, only jobs started from the current project are shown. To show other jobs, deselect Show jobs from current project only.

When a monitored job ends, the results are incorporated into the project according to the settings used to launch the job. If a job that is not currently being monitored ends, you can select it in the Monitor panel and click Monitor to incorporate the results. Monitored jobs are incorporated only if they are part of the current project. You can monitor jobs that are not part of the current project, but their results are not incorporated. To add their results to a project, you must open the project and import the results.

Further information on job control, including configuring your site, monitoring jobs, running jobs, and job incorporation, can be found in the [Job Control Guide](#) and the [Installation Guide](#).

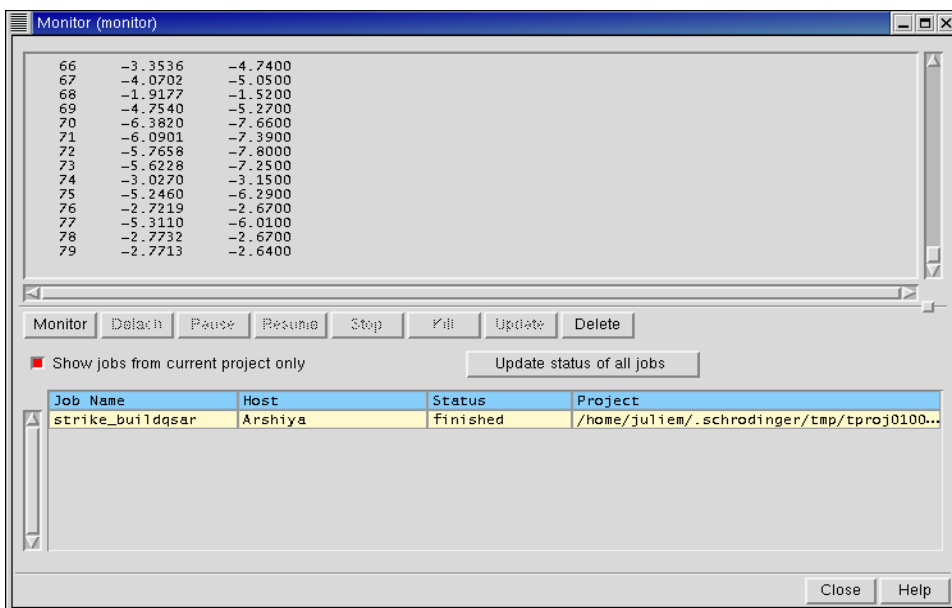


Figure 2.7. The Monitor panel.

2.11 Getting Help

Maestro comes with automatic, context-sensitive help (Auto-Help), Balloon Help (tooltips), an online help facility, and a user manual. To get help, follow the steps below:

- Check the Auto-Help text box at the bottom of the main window. If help is available for the task you are performing, it is automatically displayed there. It describes what actions are needed to perform the task.
- If your question concerns a GUI element, such as a button or option, there may be Balloon Help for the item. Pause the cursor over the element. If the Balloon Help does not appear, check that Show Balloon Help is selected in the Help menu of the main window. If there is Balloon Help for the element, it appears within a few seconds.
- If you do not find the help you need using either of the steps above, click the Help button in the lower right corner of the appropriate panel. The Help panel is displayed with a relevant help topic.
- For help with a concept or action not associated with a panel, open the Help panel from the Help menu or press CTRL+H.

If you do not find the information you need in the Maestro help system, check the following sources:

- The *Maestro User Manual*
- The Frequently Asked Questions page on the Schrödinger [Support Center](#).

You can also contact Schrödinger by e-mail or phone for help:

- E-mail: help@schrodinger.com
- Phone: (503) 299-1150

2.12 Ending a Maestro Session

To end a Maestro session, choose Quit from the Maestro menu. To save a log file with a record of all operations performed in the current session, click Quit, save log file in the Quit panel. This information can be useful to Schrödinger support staff when responding to any problem you report.

Strike Tutorial

This chapter is designed to help you become familiar with the functionality of Strike 1.5. Once you have worked through these exercises, you will have an understanding of the basic Strike features.

The Strike workflow for QSAR model generation/validation generally consists of three steps: data preparation, model generation and validation, and model application. The Strike workflow for similarity analysis using molecular properties also consists of three steps: data preparation, similarity calculation, and application of calculated similarities. For similarity analysis using two-dimensional structures (atom-pair similarity), two steps are required: the similarity calculation and application of calculated similarities. These steps will be illustrated in the tutorial exercises, which demonstrate how to do the following:

- Generate or import molecular data into Maestro for use by Strike
- Generate, validate, and apply QSAR/QSPR models
- Perform similarity analysis

Three tutorial examples are provided to demonstrate Strike workflows:

- Generating and testing a QSPR model for estimating aqueous solubility using a small number of molecular properties
- Developing a QSAR model for predicting activities of folate-based thymidylate synthase ligands
- Calculating similarities using 2-dimensional structures and molecular properties, and with these similarities extracting known actives for thermolysin from a ligand dataset.

To perform these exercises, you must have access to an installed version of Maestro 7.5 and Strike 1.5. For installation instructions, see the [Installation Guide](#).

3.1 Creating a Working Directory

The files required for the tutorial are contained in the directory `$SCHRODINGER/maestro-vversion/strike/tutorial`, in three subdirectories. They will need to be copied to a more convenient location, your local working directory.

1. Change to a directory in which you have write permission.

2. Create a new directory by entering the command:

```
mkdir workdir
```

3. Copy the Strike tutorial directories and files to your working directory:

```
cd workdir  
cp -r $SCHRODINGER/maestro-vversion/strike/tutorial/* .
```

You now have working copies of the necessary files. In the following chapters, references to tutorial files in the `qsar`, `qspr`, and `simil` directories are to the files and directories in your working directory.

3.2 Generating and Testing a QSPR Model for Aqueous Solubility

The aqueous solubility of organic molecules plays a key role in ADME processes, especially absorption, distribution, and excretion. To experimentally measure accurate aqueous solubilities ($\log S$) is difficult and requires a synthesis of the compound of interest. Because of this, a number of *in silico* approaches have been developed to estimate this key molecular property, including fragment-based approaches, linear models, and non-linear models. QikProp, Schrödinger's molecular property predictor which estimates 44 molecular properties, uses a linear method for estimating $\log S$.

The QikProp model, as with all linear or non-linear models, was fit to a finite set of compounds. When examining molecules outside the chemical space used in the fitting process, high accuracy in $\log S$ predictions might not be obtained. Consequently it may be desirable to generate local QSPR (quantitative structure-property relationship) models relevant to the compounds of interest. This tutorial provides an example of generating a local model for $\log S$ prediction using only molecular properties.

3.2.1 Starting Maestro

Now that you have a working directory tree, you can start Maestro. If you start Maestro from a particular directory, it automatically becomes your Maestro working directory. The working directory for this tutorial will be `workdir/qspr`.

When you use Maestro to launch Strike jobs, Maestro writes job output to the directory specified in the Directory folder of the Preferences panel. By default, the directory Maestro writes files to is the directory from which you launched Maestro, but you can change this directory from the Preferences panel.

If you are starting a new Maestro session:

1. Change to your `qspr` directory
2. Start Maestro by entering the following command:

```
$SCHRODINGER/maestro &
```

The Maestro main window is displayed.

3. From the Display menu, select Command Input Area.

A text box labeled Commands is added above the autohelp text area.

If you are already in a Maestro session:

1. If there is an open project (not part of this tutorial), choose Close from the Project menu.
2. In the Display menu, ensure that Command Input Area is selected.

The Maestro main window includes the command input area, as shown in [Figure 2.1](#).

3. In the Commands text box, enter the appropriate `cd` command to change to the directory `workdir/qspr`.

3.2.2 Importing Data

1. Click the Import structures button on the toolbar.



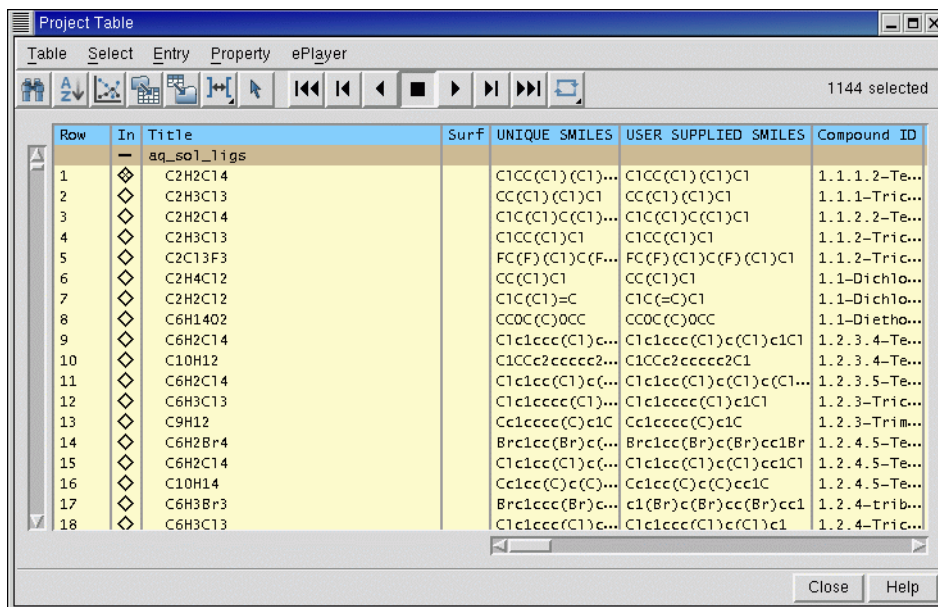
2. In the Import panel, select the Maestro-format structure file `aq_sol_ligs.mae`.

This file contains 1144 molecules for which experimental measurements of logS have been taken, as well as a set of calculated properties for each molecule.

3. Ensure that Import all structures is selected, and that the Include in Workspace option selected is First Imported Structure.
4. Click Import.

The 1144 molecular structures in the file are imported. The import operation may take a minute to finish. When it has finished, the first structure in the file is displayed in the Workspace.

5. Close the Import panel.



Row	In	Title	Surf	UNIQUE SMILES	USER SUPPLIED SMILES	Compound ID
1	◇	aq_sol_ligs				
2	◇	C2H2C14		C1CC(C1)(C1)...	C1CC(C1)(C1)C1	1.1.1.2-Te...
3	◇	C2H3C13		CC(C1)(C1)C1	CC(C1)(C1)C1	1.1.1-Tric...
4	◇	C2H2C14		C1C(C1)C(C1)...	C1C(C1)C(C1)C1	1.1.2.2-Te...
5	◇	C2H3C13		C1CC(C1)C1	C1CC(C1)C1	1.1.2-Tric...
6	◇	C2C13F3		FC(F)(C1)C(F)...	FC(F)(C1)C(F)(C1)C1	1.1.2-Tric...
7	◇	C2H4C12		CC(C1)C1	CC(C1)C1	1.1-Dichlo...
8	◇	C2H2C12		C1C(C1)=C	C1C(=C)C1	1.1-Dichlo...
9	◇	C6H14O2		CCOC(C)OCC	CCOC(C)OCC	1.1-Dietho...
10	◇	C6H2C14		C1c1ccc(C1)c...	C1c1ccc(C1)c(C1)c1C1	1.2.3.4-Te...
11	◇	C10H12		C1CCc2ccccc2...	C1CCc2ccccc2C1	1.2.3.4-Te...
12	◇	C6H2C14		C1c1ccc(C1)c...	C1c1ccc(C1)c(C1)c(C1)...	1.2.3.5-Te...
13	◇	C6H3C13		C1c1cccc(C1)...	C1c1cccc(C1)c1C1	1.2.3-Trim...
14	◇	C9H12		Cc1cccc(C)c1C	Cc1cccc(C)c1C	1.2.3-Trim...
15	◇	C6H2Br4		BrC1cc(Br)c...	BrC1cc(Br)c(Br)cc1Br	1.2.4.5-Te...
16	◇	C6H2C14		C1c1ccc(C1)c...	C1c1ccc(C1)c(C1)cc1C1	1.2.4.5-Te...
17	◇	C10H14		Cc1ccc(C)c(C)...	Cc1ccc(C)c(C)cc1C	1.2.4.5-Te...
18	◇	C6H3Br3		BrC1ccc(Br)c...	c1(Br)c(Br)cc(Br)cc1	1.2.4-trib...
19	◇	C6H3C13		C1c1ccc(C1)c...	C1c1ccc(C1)c(C1)c1	1.2.4-Trim...

Figure 3.1. The Project Table after importing 1144 structures

6. Click the Open/Close project table button on the toolbar.



The Project Table panel opens.

As shown in [Figure 3.1](#), each structure in the imported file is now an entry in the Project Table, represented by a row. The selected entries counter in the upper right corner of the panel reads 1144 selected. A long series of columns displays a number of molecular properties, or *descriptors*, which were calculated in advance for each entry. All but the first four columns (Row, In, Title, and Surf) can be scrolled into or out of view. The Project Table panel has been resized for this figure.

Strike does not generate descriptors. The descriptors in the Project Table come from three sources:

- Most of the descriptors in the table were determined by QikProp, a program distributed by Schrödinger that generates a widely applicable set of molecular properties. For more information on QikProp, see the [QikProp User Manual](#).
- A few descriptors were obtained from the ligparse utility (\$SCHRODINGER/utilities/ligparse), including the Aromatic proportion and the Non-carbon propor-

tion. The aromatic proportion is the fractional percent of heavy atoms that are aromatic while the non-carbon proportion is the fractional percent of heavy atoms that are not carbon.

- Also included are experimentally determined logS values in the measured log(solubility:mol/L) descriptor.

3.2.3 Preparing Test and Training Sets

The next step is to separate the 1144 molecules into two sets, a test set and a training set, using a random selection method that is part of the Project Table facility.

1. Choose Random from the Select menu of the Project Table panel

The Random Selection dialog box opens.

2. Ensure that the value in the Randomly select n % of entries text box is 50, the default.

By default, the random set is chosen from only the selected entries. When the structure file was imported, all entries in the project table were selected, but this may not always be the case.

3. Change the Select from option from Selected entries to All entries and click Select.

After a moment, the Project Table is redisplayed with random entries selected. The selected entries counter in the upper right corner of the panel now reads 572 selected.

To keep track of the newly selected entries, which will be used as the training set, add a column to the Project Table that labels the currently selected molecules:

1. Select Add from the Property menu to open the Add Property panel.
2. In the Name text box, type Population.
3. Choose String from the Type option menu.
4. In the Initial value text box, type training. Click Add.

A column is added to the Project Table to the right of QPlogKhsa, as shown in [Figure 3.2](#). Under the column header Population, only the currently selected entries have a value of training.

Because the random selection generator is machine-dependent, your training set is unlikely to be a precise match to that shown in [Figure 3.2](#), and therefore your results could differ from those shown in this document. Other results will also differ slightly because of differences in the random selections made.

The data has now been prepared. In the next section, it will be used to generate a model.

Row	In	Title	Surf	gKp	IP(eV)	EA(eV)	#metab	QPlogKhsa	Population
		aq_sol_ligs							
1	◇	C2H2C14		9500	00e+00	00e+00	0	-0.208521	training
2	◇	C2H3C13		9500	00e+00	00e+00	0	-0.330770	training
3	◇	C2H2C14		9500	00e+00	00e+00	0	-0.211076	training
4	◇	C2H3C13		9500	00e+00	00e+00	0	-0.326973	
5	◇	C2C13F3		9500	00e+00	00e+00	0	-0.205726	
6	◇	C2H4C12		9500	00e+00	00e+00	0	-0.459116	training
7	◇	C2H2C12		1544	00e+00	00e+00	0	-0.520913	
8	◇	C6H14O2		4834	00e+00	00e+00	0	-0.958609	training
9	◇	C6H2C14		8629	00e+00	00e+00	0	0.150132	
10	◇	C10H12		2750	00e+00	00e+00	2	0.180045	
11	◇	C6H2C14		5474	00e+00	00e+00	0	0.164577	
12	◇	C6H3C13		8257	00e+00	00e+00	0	0.025543	
13	◇	C9H12		5430	00e+00	00e+00	3	0.126947	
14	◇	C6H2Br4		1202	00e+00	00e+00	0	0.259668	
15	◇	C6H2C14		5441	00e+00	00e+00	0	0.164537	training
16	◇	C10H14		3512	00e+00	00e+00	4	0.311830	training
17	◇	C6H3Br3		0700	00e+00	00e+00	0	0.113024	training
18	◇	C6H3C13		5047	00e+00	00e+00	0	0.039972	training

Figure 3.2. The Project Table with a randomly selected training set

3.2.4 Building a Partial Least Squares Model

It is known from the general solubility equation that a relationship exists between a compound's aqueous solubility and its logP and melting points. We will use this idea in generating our model by including the logP estimate from QikProp along with a handful of molecular properties chosen to fulfill the role of the melting point.

Your first model will use the Partial Least Squares (PLS) method, which is described briefly in [Chapter 6](#). Linear equations are generated that describe the relationship between a group of factors (derived from a set of independent descriptors) and a dependent descriptor (the predicted property). The goal of PLS is to find factors that explain the variance in both the independent and the dependent descriptors.

1. Choose Build QSAR Model from the Strike submenu of the Applications menu.

The Build QSAR Model panel opens. As shown in [Figure 3.3](#), the input counter under the panel title bar reads Input is 572 entries currently selected in the Project Table.

2. Ensure that the Regression method selected is Partial Least Squares.

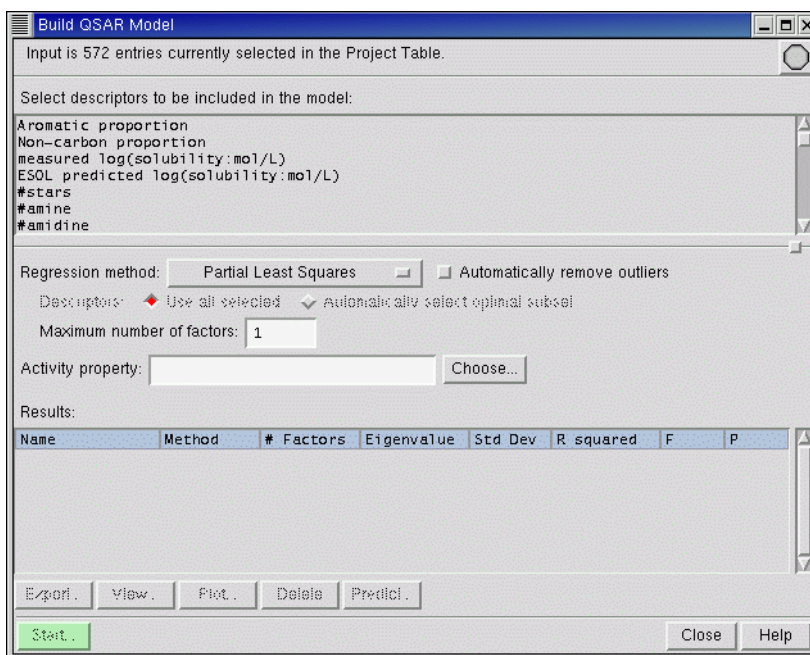


Figure 3.3. The Build QSAR Model panel

3. Under Select descriptors to be included in the model, control-click on the following:

- Aromatic proportion
- #rotor
- volume
- QPlogPo/w

The descriptor count is displayed: (4 currently selected) These four descriptors will be the independent variables.

4. Ensure that Automatically remove outliers is deselected (the default).

5. Enter 4 as the Maximum number of factors.

The Maximum number of factors should be less than or equal to the number of independent variables. If the Maximum number of factors is greater than the number of independent variables, Strike will automatically report the maximum number of factors extracted from the data, which is generally equal to the number of independent descriptors.

6. Select the Activity property (the dependent variable to be fit) by clicking Choose and selecting measured log(solubility:mol/L).

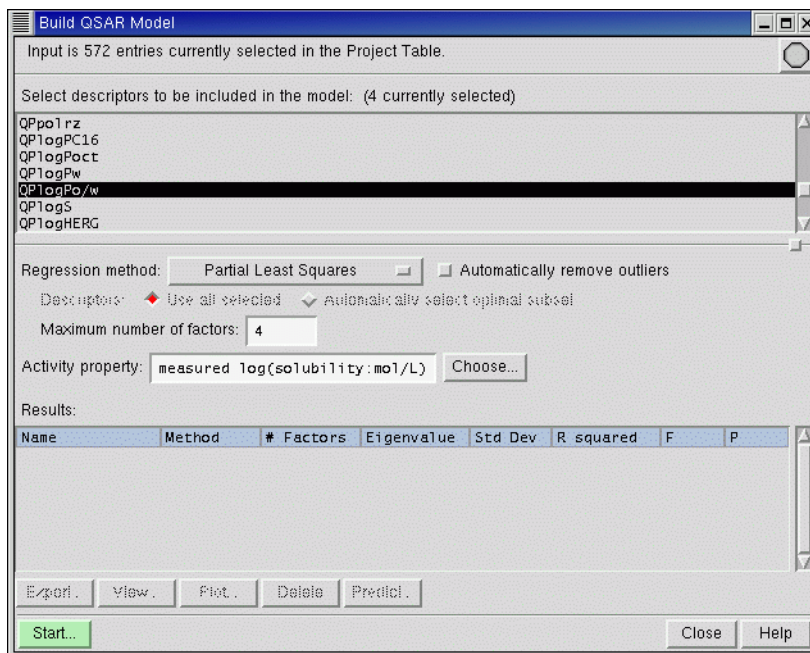


Figure 3.4. The Build QSAR Model *panel settings for the PLS model*

These settings mean that the model will attempt to correlate the number of rotatable bonds (#rotor), fractional aromatic proportion, molecular volume and logP (QPlogPo/w) to experimentally measured aqueous solubilities (measured log(solubility: mol/L)).

7. Click Start.

The Statistics / Build QSAR - Start dialog box opens.

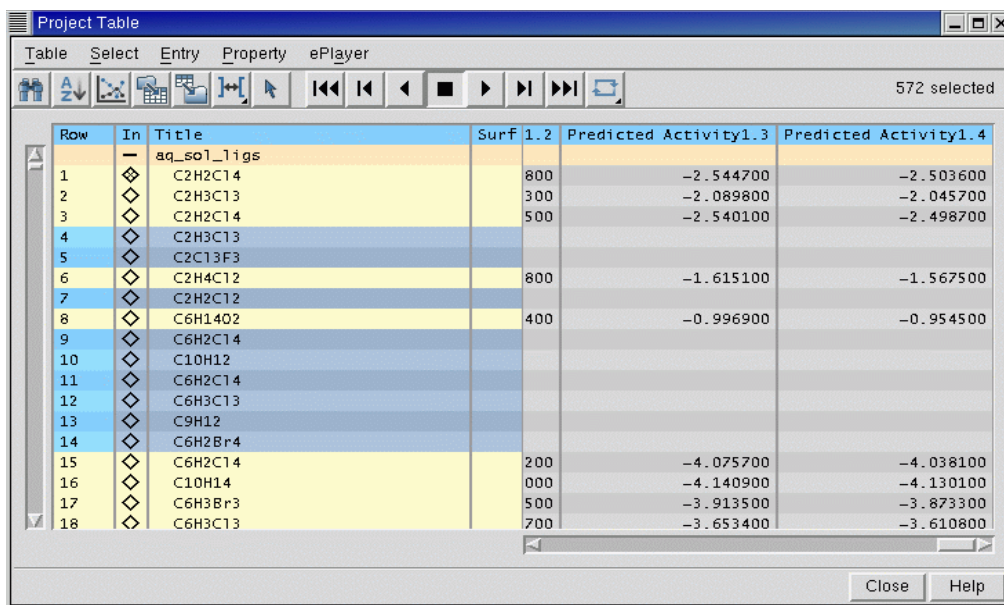
8. Select the Job options you want, then click Start to begin the calculation.

The Monitor panel opens as the job begins to run.

The job takes only a few moments to finish. When the model has been generated, the results are incorporated into the Project Table, shown in [Figure 3.5](#).

3.2.5 Examining PLS Model-Building Results

In the Project Table there are four new columns, headed Predicted Activity $X.Y$, where X represents the model and Y the number of factors used in the prediction. This is the first model built in this project, so $X = 1$, and a maximum of 4 factors were used, so $Y = 1, 2, 3, \text{ or } 4$. The values in each column are the predicted values of logS generated using Y factors.



The screenshot shows a software window titled "Project Table" with a menu bar (Table, Select, Entry, Property, ePlayer) and a toolbar. The main area displays a table with 18 rows of data. The first row is a header, and the subsequent rows list chemical structures (In, Title) along with their predicted activities (Surf, 1.2, Predicted Activity1.3, Predicted Activity1.4). The table is currently showing 18 rows, with the first row being a header and the rest being data rows. The table is sorted by Predicted Activity1.3 in descending order. The last row is highlighted in yellow.

Row	In	Title	Surf	1.2	Predicted Activity1.3	Predicted Activity1.4
1	◇	C2H2C14	800		-2.544700	-2.503600
2	◇	C2H3C13	300		-2.089800	-2.045700
3	◇	C2H2C14	500		-2.540100	-2.498700
4	◇	C2H3C13				
5	◇	C2C13F3				
6	◇	C2H4C12	800		-1.615100	-1.567500
7	◇	C2H2C12				
8	◇	C6H14O2	400		-0.996900	-0.954500
9	◇	C6H2C14				
10	◇	C10H12				
11	◇	C6H2C14				
12	◇	C6H3C13				
13	◇	C9H12				
14	◇	C6H2Br4				
15	◇	C6H2C14	200		-4.075700	-4.038100
16	◇	C10H14	000		-4.140900	-4.130100
17	◇	C6H3Br3	500		-3.913500	-3.873300
18	◇	C6H3C13	700		-3.653400	-3.610800

Figure 3.5. The Project Table with training-set predicted activities

In this document, the particular set of diagnostic statistics generated by model X with Y factors is called a *predictor*. Four predictors are listed in the Results table of the Build QSAR Model panel after the model-building job has finished, as shown in Figure 3.6. Along with the Name in the format X.Y, the Method, and the number of factors (# Factors), statistical information is given for an immediate appraisal of the predictors: standard deviation, R-squared, F-value, and P-factor.

The five buttons below the Results table are now available. When a predictor is selected, the buttons can be used to perform the following tasks. Some tasks affect only the selected predictor; others operate on the model as a whole, including any other predictors belonging to the model:

Export	Export the model for use in another project
View	View the output file for the model-building job that generated the predictor
Plot	Plot the predicted versus experimental results for the selected predictor
Delete	Delete the model that generated the predictor
Predict	Make further predictions using the selected predictor

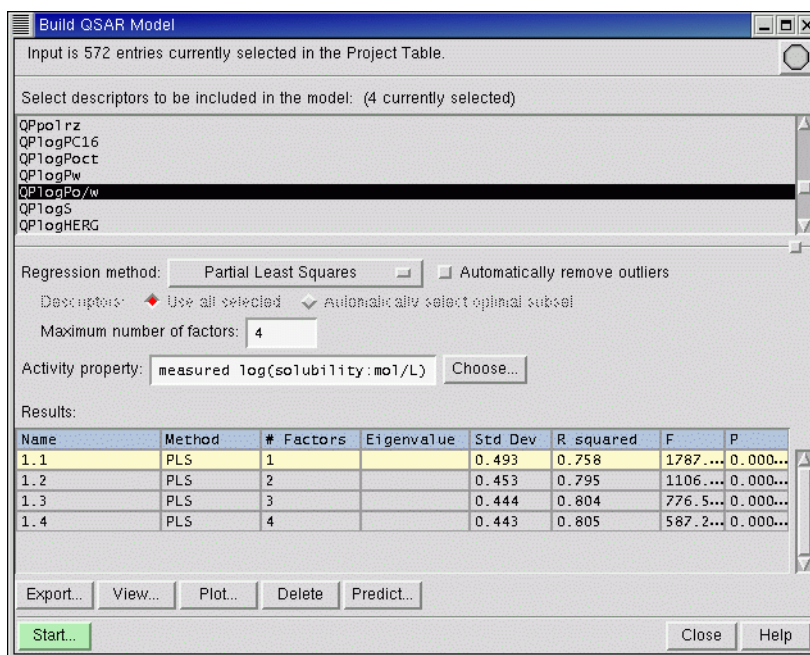


Figure 3.6. The Build QSAR Model panel with a four-predictor PLS model

Examine the 4-factor predictor (1.4) by plotting its predictions for the training set versus measured log(solubility: mol/L):

1. Select the row named 1.4 in the Results table and click Plot.

After a few moments, the Plot XY panel appears, showing measured logS values plotted against predicted values.

By default, the Plot XY facility automatically sets the range for the X and Y axes of the plot. However, it is easier to spot outlying points when the X and Y axes share a common scale. If instead one spans a broader range of values than the other, this can be adjusted in the Plot Settings panel.

2. Choose Plot Settings from the Settings menu.

The Plot Settings panel opens.

3. Click the tab (X Axis or Y Axis) for the axis with the smaller range.
4. In the X Axis or Y Axis folder, click the Edit button.

The Edit Axis dialog box opens.

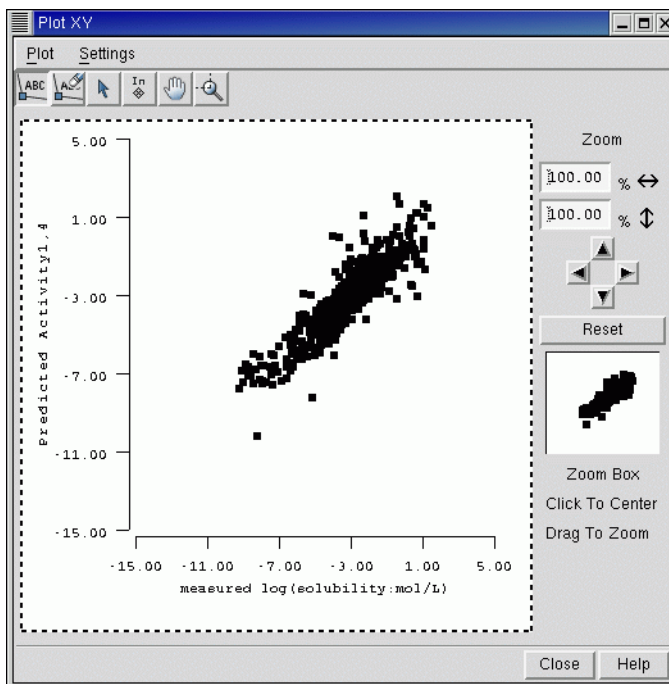


Figure 3.7. Plot of predicted vs. measured logS for training set

5. Deselect Auto range.

The Minimum and Maximum text boxes are now available.

6. Adjust the values to match the axis with the larger range, and click Edit.

The dialog box closes.

7. Close the Plot Settings panel.

The plot is redrawn to a common scale, as in the example in [Figure 3.7](#). Because the training set is selected randomly, some plot details will differ from yours.

For most of the molecules in this training set, the generated model does a good job of reproducing experimental logS values, as can be seen in [Figure 3.7](#). Next you will examine some of the cases where the model generates less-accurate predictions:

1. In the Plot XY toolbar, click the Include project entries button.



2. Pick one of the data points for which the model is most in error.

The corresponding molecule is now included in the Workspace.

3. Pick other data points where the model is in error and look at their molecular structures in the Workspace.

The particular molecules you view will depend on the training set, but you should find that the model does less well for molecules that contain long alkyl chains, sugar-like molecules, and a few fused-ring heterocycles.

For information about other features of the Plot XY panel, such as the Zoom box, Pan, and point labeling tools, click the Help button.

4. When you have finished working with the panel, click Close.

More information about the model and the predictors is given in the output file of the model-building job, *jobname.out*, which can be examined using the View button:

1. In the Build QSAR Model panel, click the View button.

The View QSAR Model dialog box opens. This dialog box displays the output file for the Strike model-building job—see [Figure 3.8](#).

2. Examine the output file, noting the following points of interest:
- The Correlation Matrix for input variables (the four independent descriptors).
 - PLS Regression Statistics, listing standard deviation (S.D.), R-squared, F-factor, and P-value by #Factors. The large F-factors and small P-values indicate this model was likely not achieved by chance and that the descriptors chosen are significant as a set.

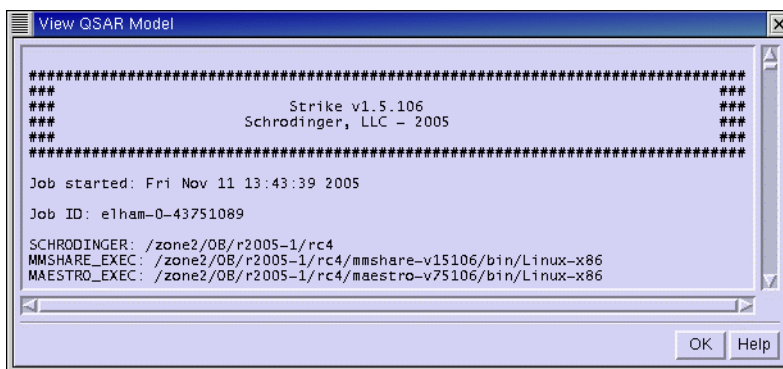


Figure 3.8. The View QSAR Model dialog box

- Cross Validation leave- N -out Results over M Cycles. Large differences between calculated q^2 and r^2 values reflect significant dependence of the model on the molecules included in the regression and in general are unfavorable.
 - T-values and coefficients.
 - Predicted values for logS at each #Factors for each of the 572 molecules in the training set.
3. Click OK to close the View QSAR Model dialog box.

If a job fails, the View button will not display the output file. Examine the output file *jobname.out* in a terminal window instead.

3.2.6 Applying the Model to the Test Set

The true test of any model is to check its predictions against a set of molecules not included during its training. The exercise performed in this section would typically be considered part of model generation and validation, but for the purposes of this tutorial, it will be used to demonstrate the model application step of the Strike workflow.

The first step is to create a test set of molecules. In this example, the test set will be those molecules in the Project Table that were not members of the training set:

1. In the Project Table, confirm that the training set is selected by examining the Population column. If so, skip to the next step.

If for any reason the training set is no longer the selected set—for example, if a single entry has been selected instead—you can restore the selection by performing these steps:

- a. Choose Only from the Select menu of the Project Table.

The Entry Selection panel opens.

- b. In the Properties list, select Population.
- c. Select the option Is defined (any value).

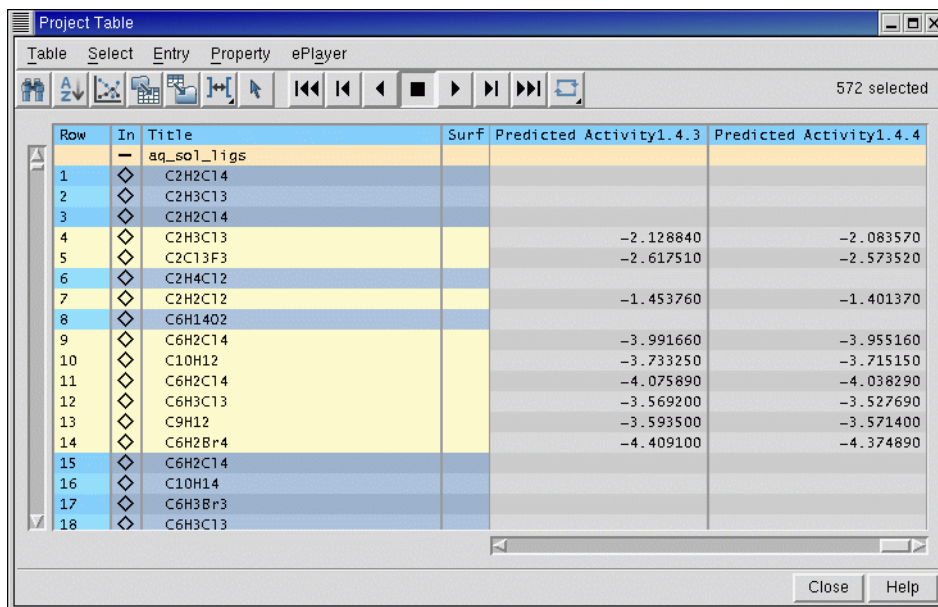
Only the training set has a defined value (training) in the Population column.

- d. Click Add, then OK.

The molecules in the training set, and only those molecules, are now selected.

2. In the Project Table, choose Invert from the Select menu.

The selected molecules are now those that were not part of the training set. This will be the test set.



Row	In	Title	Surf	Predicted Activity1.4.3	Predicted Activity1.4.4
		aq_sol_ligs			
1	◇	C2H2C14			
2	◇	C2H3C13			
3	◇	C2H2C14			
4	◇	C2H3C13		-2.128840	-2.083570
5	◇	C2C13F3		-2.617510	-2.573520
6	◇	C2H4C12			
7	◇	C2H2C12		-1.453760	-1.401370
8	◇	C6H14O2			
9	◇	C6H2C14		-3.991660	-3.955160
10	◇	C10H12		-3.733250	-3.715150
11	◇	C6H2C14		-4.075890	-4.038290
12	◇	C6H3C13		-3.569200	-3.527690
13	◇	C9H12		-3.593500	-3.571400
14	◇	C6H2Br4		-4.409100	-4.374890
15	◇	C6H2C14			
16	◇	C10H14			
17	◇	C6H3Br3			
18	◇	C6H3C13			

Figure 3.9. The Project Table with test-set predicted activities.

Run a prediction job on the test set molecules:

1. In the main window, choose Predict from the Strike submenu of the Applications menu.

The Predict based on QSAR model panel opens. If it was open, the Build QSAR Model panel closes.

In the Predict panel, the four predictors previously generated are listed in the Select model to use for prediction table.

2. Select the model with 4 in the # Factors column.
3. Click the Start button to open the Statistics / Predict - Start dialog box. Change the Host and Username if necessary before clicking Start to launch the `strike_predict` job.

The Monitor panel appears. The job takes a few seconds to run.

4. When the job is finished, view the Project Table.

There are four new columns to the right of the table: Predicted Activity1.4.*N*, where *N*=1, 2, 3, or 4. These columns hold predicted logS values for the test set, as shown in Figure 3.9.

Using the data in the Predicted Activity1.4.4 column of the Project Table as the predicted logS, plot the predicted logS versus measured log(solubility:mol/L) for the test set molecules:

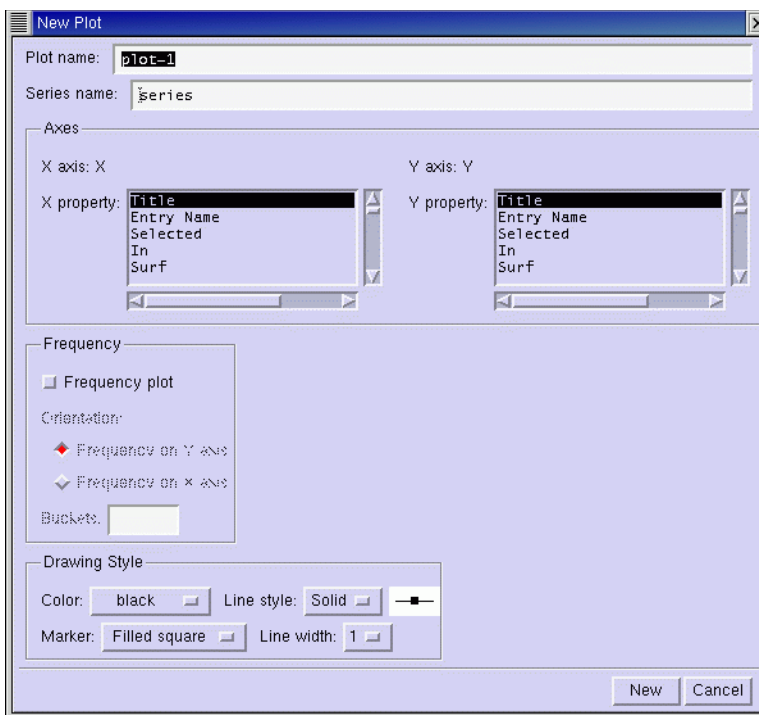


Figure 3.10. The New Plot dialog box with default settings

1. Open the Plot XY panel by clicking the Plot button in the Project Table toolbar



or selecting Plot from the Table menu.

The Plot XY panel opens with the most recent plot displayed. This plot, like all the plots created so far, uses the training set data.

2. Choose New Plot from the Plot menu.

The New Plot dialog box opens.

3. Change the following settings:

- a. In the Plot name text box, type pls144test.
- b. In the Series name text box, type tutorial.
- c. In the Axes section, select measured log(solubility:mol/L) from the X property list.

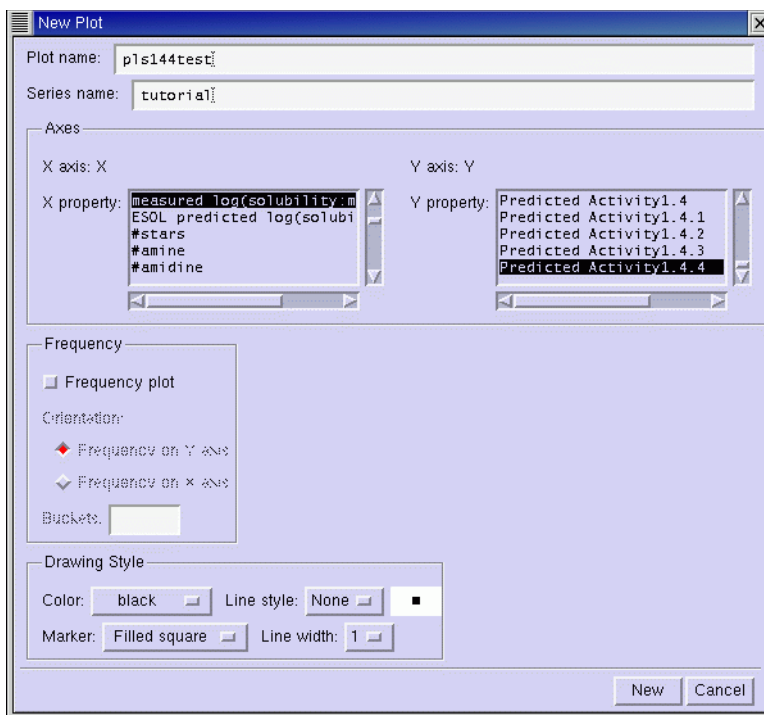


Figure 3.11. The New Plot dialog box with settings for `pls144test` plot

- d. Select Predicted Activity1.4.4 from the Y property list.
- e. In the Drawing Style section, select None from the Line style list.

These settings are shown in Figure 3.11.

4. Click New to close the dialog box and display the plot `pls144test` in the Plot XY panel.

The new plot appears to the right of the training set plot. By default, the newly added plot is selected. The old plot can now be deleted.

5. Select the old plot by clicking within it.

The dashed line that previously enclosed the new plot is transferred to the old plot, indicating that it is now the selected plot.

6. From the Plot menu, choose Delete Selected Plots.

The selected plot (the old plot) is deleted, and the new plot occupies the full width of the panel.

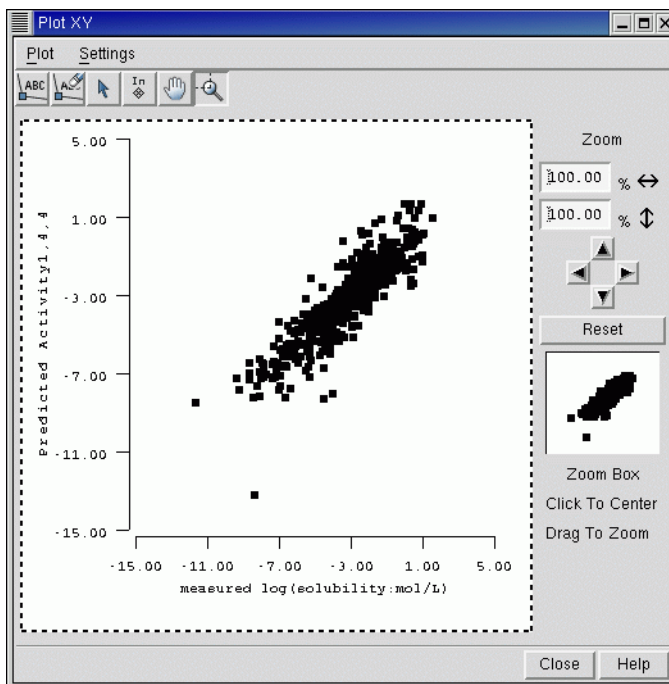


Figure 3.12. Plot of predicted vs. measured logS for the test set

Now adjust the range of values on one of the axes to match the other.

- Click in the new plot to select it, then choose Plot Settings from the Settings menu and repeat the steps that produced [Figure 3.7 on page 39](#).

The plot is displayed, resembling that shown in [Figure 3.12](#). The good agreement between calculated and experimental values found for the training set remains quite good for the test set.

- Select the **In** button on the Plot XY toolbar and click on one or more of the less well-treated members of the test set.

As they appear in the Workspace, note that many are fused heterocycles, sugar-like molecules, or molecules with large aliphatic chains, as was observed in the training set.

3.2.7 Calculating Univariate and Bivariate Statistics

The Strike statistics script calculates univariate and bivariate statistics of selected descriptors for the set of entries currently selected in the Project Table. The Strike Univariate and Bivariate Statistics panel allows you to select from a list of the descriptors found in the Project Table.

When you have run a statistics job, the results are displayed in the dialog box, from which information can be copied and pasted to an open file as reference material or for printing.

1. From the Strike submenu of the Applications menu in the main window, choose Statistics.
2. Select `Aromatic_proportion` from the list under Select one or two descriptors.

The selected descriptor is highlighted in yellow.

This will be a univariate statistics calculation, so the only input needed is the single descriptor you have selected.

3. Click Start to launch the job under the default name, `strikeStats`.

After a moment, the job results appear in the Results text area, as shown in [Figure 3.13](#). These univariate statistics describe the range and variance of the descriptor values and the shape of the distribution for the test set of molecules (the currently selected entries in the Project Table). See [Chapter 6](#) for definitions of statistics terms.

Now set up a bivariate statistics calculation.

4. From the descriptor list, select `measured_log(solubility:mol/L)`, then control-click on `#rotor`.

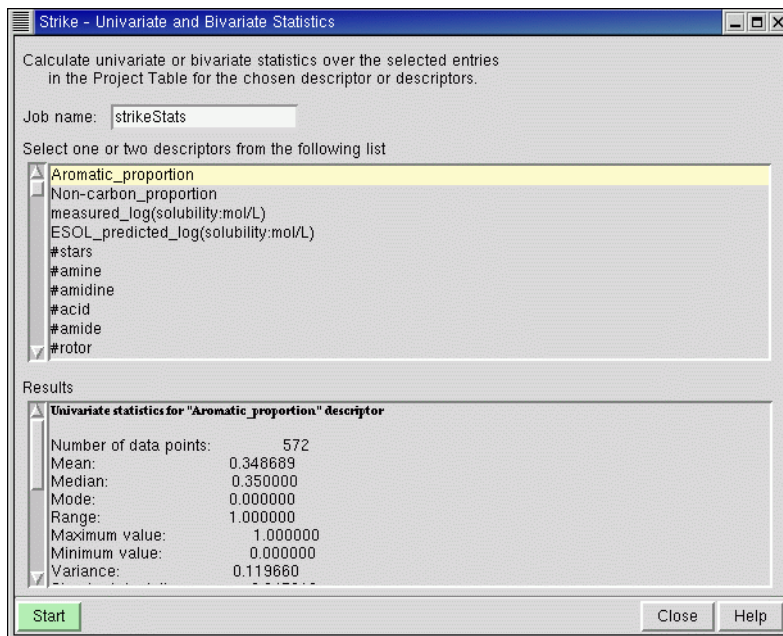


Figure 3.13. The Strike Statistics dialog box with univariate statistics

5. Click Start.

After a few moments, the new results are appended to the Results table, and you will need to scroll down to view them. They include a small set of bivariate statistics for the pair of descriptors, followed by the univariate statistics for each, as shown in [Figure 3.14](#). See [Chapter 6](#) for definitions of statistics terms.

The dialog box obtains its descriptor information from the Project Table. If you subsequently add or remove properties (descriptors) from the Project Table, you need to close the statistics dialog box and then reopen it to capture the new information.

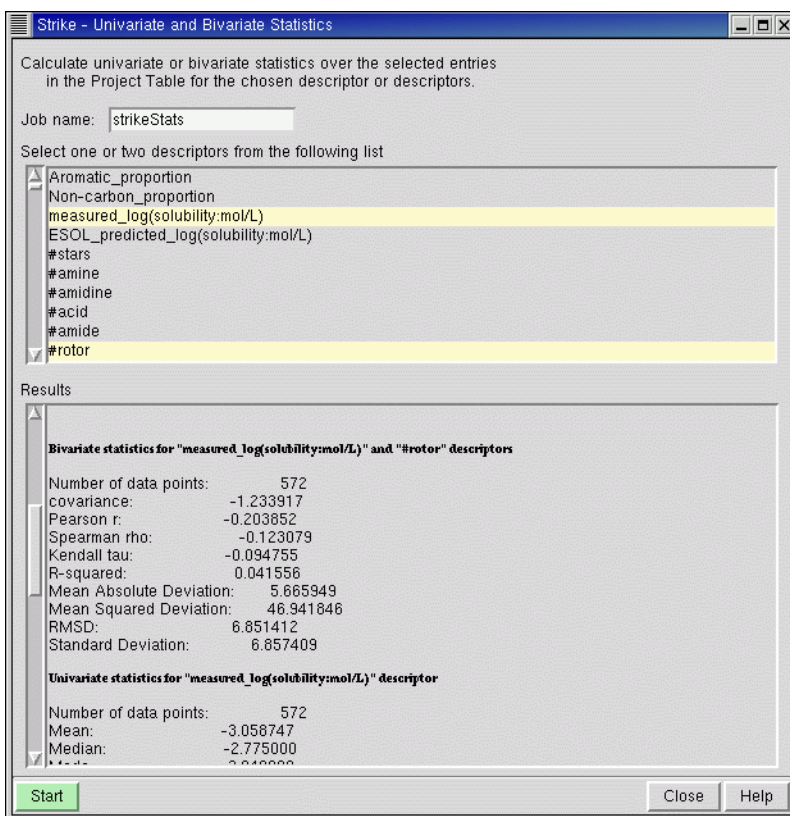


Figure 3.14. The Strike Statistics dialog box with bivariate statistics

3.2.8 Model-Building Using Principal Component Analysis

In this exercise, you will generate a model using one of the other alternative regression methods available in Strike, Principal Component Analysis.

1. In the Project Table, the test set is currently selected. Choose Invert from the Select menu to select the training set.
2. Open the Build QSAR Model panel.

The Predict panel closes. The Build QSAR Model panel retains the selected descriptors, number of factors, and regression method used to generate the previous model. The four predictors generated by the PLS model-building job remain in the Results table.

3. Choose Principal Component Analysis from the Regression method option menu and click Start.
4. In the Start dialog box, change the job name to `strike_buildqsar_pca` and click Start to launch the calculation.

When the job has finished, the new model will be added to the Results table as a set of predictors with # Factors equal to 1, 2, 3, and 4, as was the PLS model. In the Eigenvalue column, a number is associated with each of the four PCA-model predictors. The eigenvalue represents the portion of the total variance accounted for by the n -factor predictor.

In the Project Table, the four new columns Predicted Activity_{2.n} are added to the table, with values only for the training set of molecules.

If you were using this PCA model in a real project, you could continue by carrying out the analysis and prediction steps that were performed for the PLS model earlier in this chapter.

The PCA method is frequently used for data reduction by retaining only those factors needed to account for most of the total variance. The variance of each of the independent variables used in the model is taken to be 1.0, and the total variance is defined as the sum of the variances of each independent variable. Typically it is sufficient to retain only those factors with an eigenvalue greater than 1.0. These are the factors that account for more of the variance than does any single original variable.

In the Results table in the Build QSAR Model panel, the n -factor predictor of a PCA model accounts for a portion of the total variance equal to the sum of the first n eigenvalues. For example, in the table shown in [Figure 3.15](#), the first eigenvalue is 1.78 and the second 1.33, while the third and fourth eigenvalues are less than 1.0. The total variance is 4.0, and the two-factor predictor is sufficient to account for $(1.78 + 1.33)/4.00 = 78\%$ of the total variance.

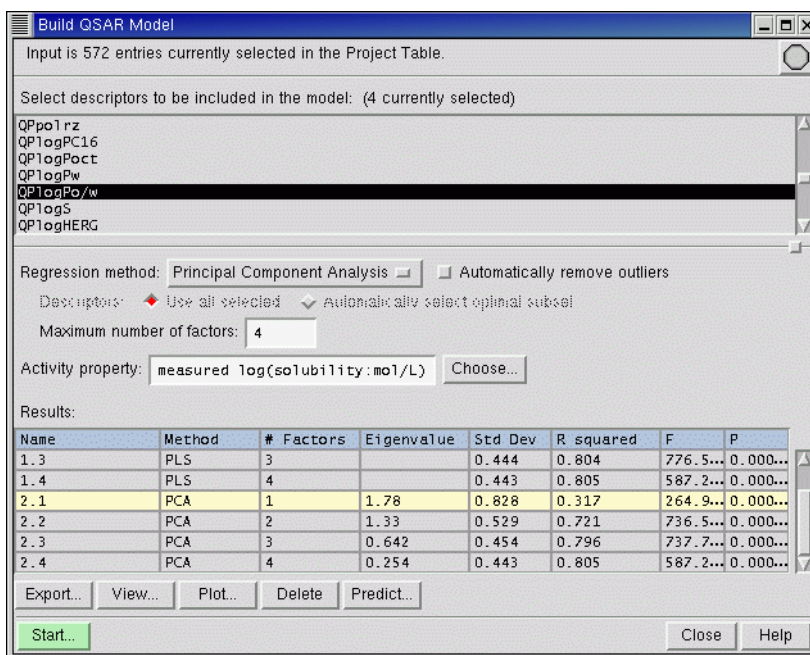


Figure 3.15. Build QSAR Model *panel with four-predictor PCA model*

3.2.9 Model-Building Using Multiple Linear Regression

The third regression method available for model-building in Strike is multiple linear regression (MLR). In this section, you will generate an MLR model that uses an algorithm to select the optimal set of descriptors for use.

1. Ensure that the training set is selected in the Project Table.
2. In the Select descriptors to be included in the model list, use control-click to add two more descriptors, mol MW and SASA.
3. Choose Multiple Linear Regression from the Regression method option menu.
4. Select the Descriptors option Automatically select optimal subset.
5. Ensure that the Size of optimal subset is 4.

These settings instruct the MLR algorithm to use the best subset of four descriptors from the six selected.

6. Click Start.

The Statistics / Build QSAR - Start dialog box opens.

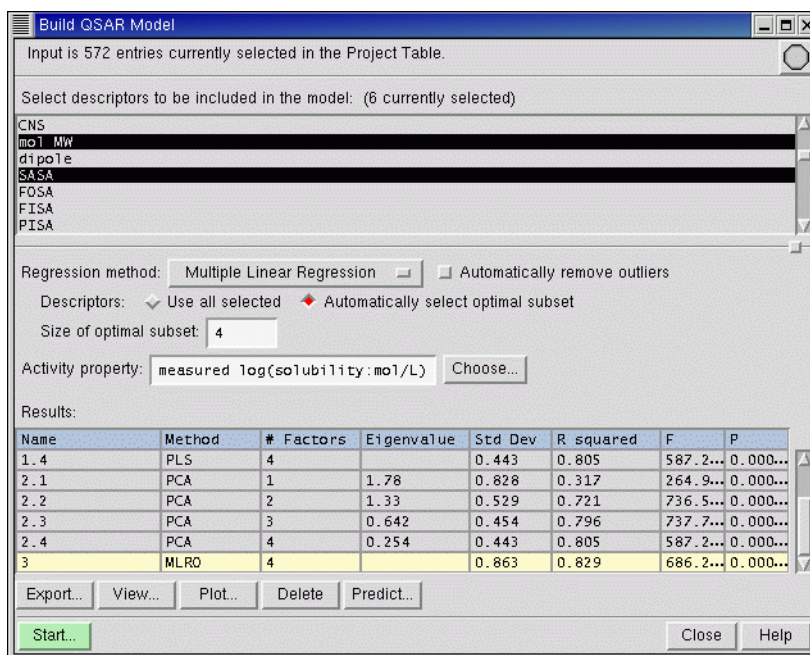


Figure 3.16. Build QSAR Model *panel with MLR model*

- Change the job name to `strike_buildqsar_mrl`.
- Select the other job options you want, and click **Start** to launch the calculation.

The job takes only a few moments to finish. When the model has been generated, it appears in the Results table in the Build QSAR Model panel as a single row with MLRO (MLR optimal subset) as the Method. See [Figure 3.16](#).

The results are also incorporated into the Project Table as the column Predicted Activity3.

Again, you could continue with analysis and prediction using this model, as you did with the PLS model.

As you have seen, Strike can be used to easily generate accurate and widely applicable QSPR models using molecular properties. This may be applied to generate local or global models for a number of properties using your own in-house data.

3.3 Calculating Atom-Pair Similarities

It is often useful to identify molecules that are “similar” in a chemically significant way to structures of interest. Strike can be used to analyze similarity in either two-dimensional structural (atom-pair connectivity) or molecular descriptor space. Using the atom-pair connectivity method has the advantage of requiring only structural (connectivity) information for a set of probe molecules and for the molecules for which calculated similarities are desired.

In this section, you will use Strike to calculate atom-pair similarities, then use those similarities to extract known actives for thermolysin from a ligand data set.

3.3.1 Changing Maestro Directories

If you are starting a new Maestro session:

1. Change to your *working-directory/simil/* directory and start Maestro by entering

```
$SCHRODINGER/maestro &
```

The Maestro main window is displayed.

If you are already in a Maestro session:

1. Choose Close from the Project menu.

If the project is a scratch project from the previous exercises, you may discard it.

2. In the Display menu, ensure that Command Input Area is selected.

The Maestro main window includes the command input area.

3. In the Commands text box, enter the appropriate `cd` command to change to the directory *working-directory/simil*.

3.3.2 Importing Active and Decoy Ligands

First we need to create a ligand database which is seeded with a subset of the known active ligands. Using the remaining active ligands as probe molecules, we will attempt to extract the seeded actives out of the database.

1. Click the Import structures button in the toolbar.



2. In the Import panel, select the Maestro-format structure file `1tmn_actives.mae`.

This file contains nine known active ligands for thermolysin.

3. Ensure that Import all structures is selected, and that the Include in Workspace option selected is First Imported Structure.
4. Click Import.

The first active ligand structure appears in the Workspace.

5. Open the Project Table.

There are nine entries, each with a Title identifying the ligand. Each row also has columns of calculated molecular properties from QikProp. These properties will not be used in this exercise, as they are not needed to generate atom-pair similarities.

6. Display each of the ligands in turn in the Workspace.

These structures show some diversity though many have peptide moieties.

Note: if the Workspace appears empty, click the Fit to screen toolbar button



to bring the ligand into view.

7. In the Import panel, select the file `d1-400mw.mae` and click Import. This file contains the Maestro-format structures of 998 decoy ligands with an average molecular weight of 400.

After a few moments, the first decoy structure appears in the Workspace and 998 new entries are added to the Project Table. As can be seen in [Figure 3.17](#), these entries also have molecular properties calculated by QikProp which will not be used in this exercise.

8. Close the Import panel.

3.3.3 Opening the Calculate Similarity Panel

1. Choose Similarity from the Strike submenu of the Applications menu.

The Calculate similarity panel opens, as shown in [Figure 3.18](#). As noted in the panel, similarity will be calculated for the entries selected in the Project Table, using the entries included in the Workspace as probe molecules.

2. Ensure that the Atom pair similarities option is selected.

Project Table

Table Select Entry Property ePlayer

998 selected

Row	In	Title	Surf	QPPMDCK	QLogKp	IP(eV)	EA(eV)	#metab	QLogKhsa
	—	1tmm_actives							
1	◇	1l1na		.439690	.136152	501985	280343	000000	-1.209945
2	◇	1th1		.820292	.728250	192592	058492	000000	0.192351
3	◇	1tlp		.366887	.944287	461018	309641	000000	-1.499553
4	◇	1tmm		.687191	.916261	318049	048195	000000	-0.204961
5	◇	3tmm		.235154	.393757	294335	036350	000000	-0.611643
6	◇	4tmm		.856473	.993145	002010	098317	000000	-0.622220
7	◇	5t1n		.290522	.981824	565976	091317	000000	-1.664245
8	◇	5tmm		.725346	.907379	517671	063740	000000	-0.859533
9	◇	6tmm		.384328	.254253	427417	060537	000000	-0.772346
	—	d1-400mw							
10	◇	18		.400572	.799400	066854	001301	000000	-0.237435
11	◇	27		.831570	.227575	945114	040910	00e+00	0.611880
12	◇	35		.849733	.810180	844566	973439	000000	0.919937
13	◇	44		.489904	.272433	072822	904458	000000	0.753041
14	◇	57		.245709	.687001	825658	538242	000000	0.687851
15	◇	76		.786293	.333201	193809	129293	000000	0.568793
16	◇	95		.518148	.139586	092597	004997	000000	0.319514
17	◇	123		.078196	.620410	695763	883990	000000	0.471178

Close Help

Figure 3.17. The Project Table with 9 active and 998 decoy ligands.

3.3.4 Seeding the Database and Designating Probes

At this point, all of the decoy ligands and none of the actives are selected in the Project Table. In this exercise, you will include three active ligands in the Workspace and add the other six active ligands to the selection to create the seeded ligand set.

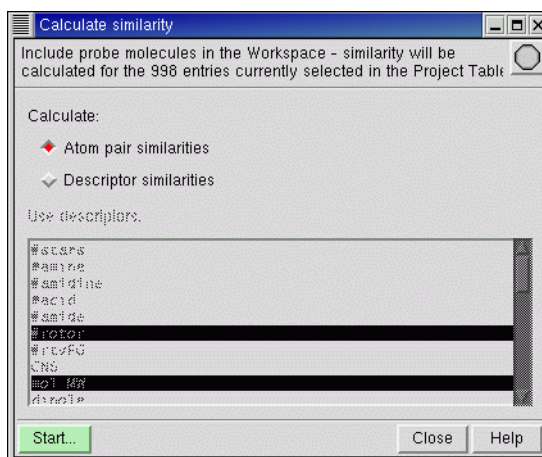


Figure 3.18. The Calculate Similarity panel

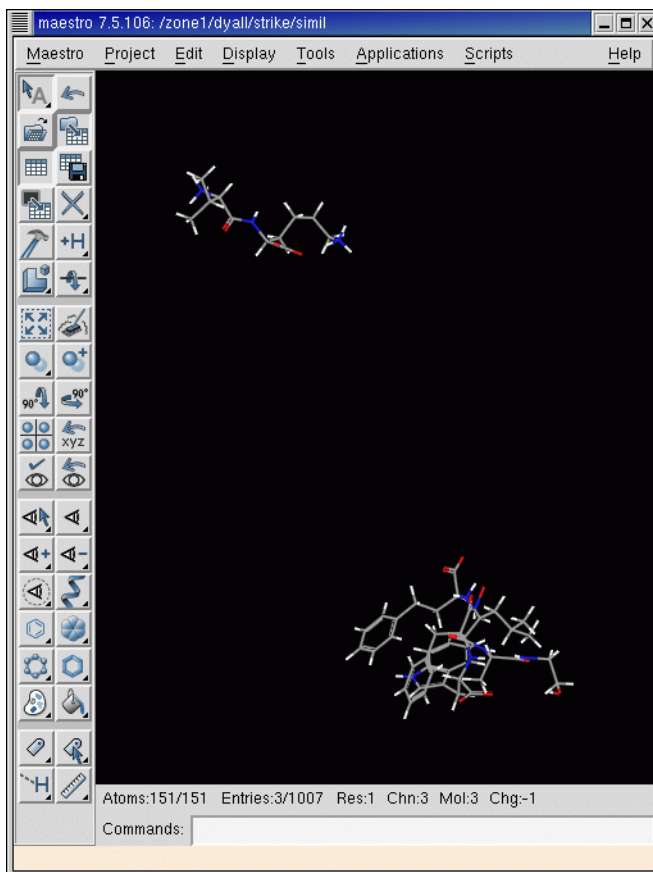


Figure 3.19. The three probe molecules included in the Workspace

1. Include the three active ligands, 11na, 1tmn, and 5tln in the Workspace

Use control-click for the second and third ligands.

These three entries are *not* selected in the Project Table.

Figure 3.19 shows the three probe molecules. You may have to click the Fit to screen toolbar button to bring them into view.

2. Add the six active ligands that are not included in the Workspace (1thl, 1tlp, 3tmn, 4tmn, 5tmn, and 6tmn) to the selected entries by control-clicking their rows.

The database for which similarities will be calculated now contains 1004 entries, of which six are known actives. The resulting Project Table is shown in Figure 3.20.

Project Table --- Scratch Project

TableSelectEntryPropertyePlayer

1004 selected

Row	In	Title	Surf	QPPMDCK	QPlogKp	IP (eV)	EA (eV)	#metab	QPlogKhsa
1tmn_actives									
1	◇	1lna		.439690	.136152	501985	280343	000000	-1.209945
2	◇	1thl		.820292	.728250	192592	058492	000000	0.192351
3	◇	1tlp		.366887	.944287	461018	309641	000000	-1.499553
4	◇	1tmn		.687191	.916261	318049	048195	000000	-0.204961
5	◇	3tmn		.235154	.393757	294335	036350	000000	-0.611643
6	◇	4tmn		.856473	.993145	002010	098317	000000	-0.622220
7	◇	5tmn		.290522	.981824	565976	091317	000000	-1.664245
8	◇	5tmn		.725346	.907379	517671	063740	000000	-0.859533
9	◇	6tmn		.384328	.254253	427417	060537	000000	-0.772346
d1-400mw									
10	◇	18		.400572	.799400	066854	001301	000000	-0.237435
11	◇	27		.831570	.227575	945114	040910	00e+00	0.611880
12	◇	35		.849733	.810180	844566	973439	000000	0.919937
13	◇	44		.489904	.272433	072822	904458	000000	0.753041
14	◇	57		.245709	.687001	825658	538242	000000	0.687851
15	◇	76		.786293	.333201	193809	129293	000000	0.568793
16	◇	95		.518148	.139586	092597	004997	000000	0.319514
17	◇	123		.078196	.620410	695763	883990	000000	0.471178

CloseHelp

Figure 3.20. The Project Table with 1004 entries selected, 3 probes included.

3. Click Start.

The Statistics / Similarity - Start dialog box opens.

4. Choose job options, then click Start to run the calculation.

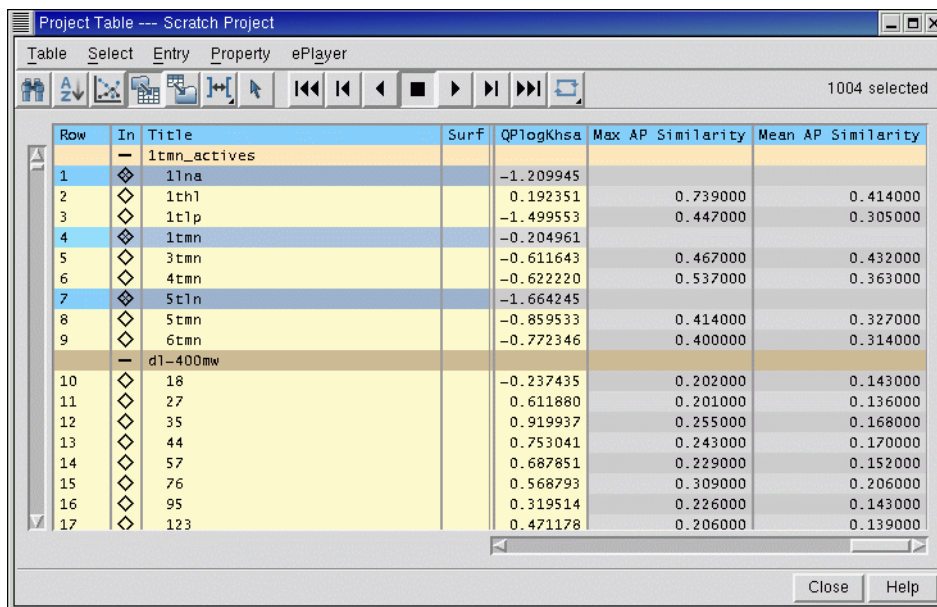
The Monitor panel appears as the job begins to run.

Two columns are added to the Project Table upon completion of the job, as shown in [Figure 3.21](#). For each entry, the Max AP Similarity is the maximum atom-pair similarity of that structure to any of the probe molecules, and the Mean AP Similarity is the mean of the atom-pair similarities to each of the probes.

By default, atom-pair similarities are calculated on a scale from 0.0 to 1.0, with 0.0 indicating no structural similarity and 1.0 indicating maximum structural similarity.

3.3.5 Applying Atom-Pair Similarity

In this exercise, you will examine how well atom-pair similarity performs in extracting the six active ligands (those not used as probes) from the set of decoy ligands. To do this, you will sort the entries in the Project Table by similarity. Project Table entries (rows) can be sorted by multiple user-specified properties (columns) called primary, secondary, and tertiary keys.



Row	In	Title	Surf	QPlogKhsa	Max AP Similarity	Mean AP Similarity
		1tmn_actives				
1	◇	1lna		-1.209945		
2	◇	1thl		0.192351	0.739000	0.414000
3	◇	1tlp		-1.499553	0.447000	0.305000
4	◇	1tmn		-0.204961		
5	◇	3tmn		-0.611643	0.467000	0.432000
6	◇	4tmn		-0.622220	0.537000	0.363000
7	◇	5tmn		-1.664245		
8	◇	5tmn		-0.859533	0.414000	0.327000
9	◇	6tmn		-0.772346	0.400000	0.314000
		dl-400mw				
10	◇	18		-0.237435	0.202000	0.143000
11	◇	27		0.611880	0.201000	0.136000
12	◇	35		0.919937	0.255000	0.168000
13	◇	44		0.753041	0.243000	0.170000
14	◇	57		0.687851	0.229000	0.152000
15	◇	76		0.568793	0.309000	0.206000
16	◇	95		0.319514	0.226000	0.143000
17	◇	123		0.471178	0.206000	0.139000

Figure 3.21. The Project Table showing atom pair similarity results.

When you imported the entries, they were grouped according to the file they were imported from. To sort the entries they must first be ungrouped.

1. Select the 1tmn_actives group, and choose Ungroup from the Entry menu.
2. Click Delete when prompted to delete the empty group.
3. Select the dl-400mw group, and choose Ungroup from the Entry menu.
4. Click Delete when prompted to delete the empty group.
5. Click the Sort button on the Project Table panel toolbar.



The Sort Project Table panel opens, as shown in [Figure 3.22](#).

6. Choose Max AP Similarity from the Primary Key list.
7. Choose Descending from the Order option menu under the Primary Key list.
8. Click Sort All Rows.

The Project Table is sorted by descending Max AP Similarity, as in [Figure 3.23](#).

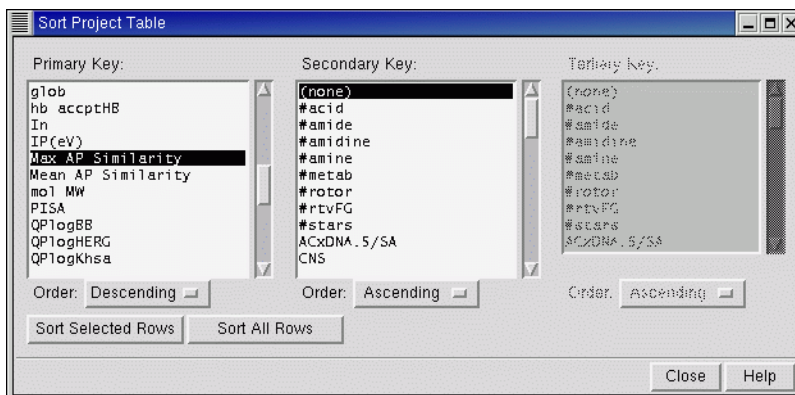


Figure 3.22. The Sort Project Table *panel*.

All six actives were found within the first 41 compounds (4.1% of the data set), and five in the first 28 compounds (2.8% of the data set). The probe molecules are at the end of the Project Table.

- As a second test, sort the Project Table again, repeating the previous steps, but this time using Mean AP Similarity as the Primary Key.

Row	In	Title	Surf	QPlogKhsa	Max AP Similarity	Mean AP Similarity
1	◇	1th1		0.192351	0.739000	0.414000
2	◇	8455		0.354198	0.546000	0.294000
3	◇	4tmn		-0.622220	0.537000	0.363000
4	◇	471565		-0.163419	0.480000	0.386000
5	◇	3tmn		-0.611643	0.467000	0.432000
6	◇	624664		-0.193329	0.464000	0.320000
7	◇	559347		0.565545	0.462000	0.306000
8	◇	430157		0.673676	0.458000	0.262000
9	◇	419868		1.246699	0.458000	0.256000
10	◇	395867		0.713440	0.455000	0.323000
11	◇	785217		0.824611	0.450000	0.336000
12	◇	1t1p		-1.499553	0.447000	0.305000
13	◇	557777		-0.371101	0.446000	0.320000
14	◇	788463		0.109787	0.444000	0.304000
15	◇	975000		0.969139	0.442000	0.255000
16	◇	723392		0.571568	0.440000	0.307000
17	◇	333967		-0.116023	0.430000	0.304000
18	◇	773364		0.904190	0.428000	0.290000
19	◇	724060		0.197419	0.425000	0.233000

Figure 3.23. The Project Table *sorted by* Max AP Similarity.

Using the mean atom-pair similarities, all six actives are found in the first 21 compounds (2.1% of the data set).

It is not surprising that the mean atom-pair similarity does a better job of extracting actives from the data set than the maximum atom-pair similarity. Because all actives are at least slightly structurally similar, their mean values are raised compared to decoy ligands, which may share common features with only one active molecule.

This example shows that Strike can be used to extract compounds similar to a set of molecules using 2D-geometry atom-pair similarities. Next, you will perform this extraction using descriptor similarity instead of atom-pair similarity.

3.4 Calculating Descriptor Similarities from Molecular Properties

Now you will use calculated molecular properties to test the ability to extract actives from the data set using descriptor similarities. The molecular properties for the thermolysin active ligands and decoy ligands were previously determined using QikProp. From this set of molecular properties, four descriptor-based similarities can be calculated: *Euclidean similarity*, *Euclidean squared similarity*, *Manhattan similarity*, and *Tanimoto similarity*. Each of these methods calculates the descriptor-space distance between two molecules as a function of their molecular properties. For a summary of each of these methods, see [Chapter 6](#).

The calculated similarities for all but Tanimoto similarity are expressed as distances on an arbitrary scale, where the smaller the value (the shorter the distance), the more similar the two molecules. High values for these quantities correspond to longer distances in descriptor space, indicating less similarity.

The Tanimoto similarity is calculated on a scale from 0.0 to 1.0 with 1.0 indicating maximum similarity and 0.0 indicating no similarity.

Like atom-pair similarity, Strike calculates descriptor similarities for a set of molecules relative to one or more probe molecules. The molecules included in the Workspace are used as probes. In this example, the probes will also be included in the test set.

1. In the Project Table, select all entries.
2. Ensure that ligands 11na, 1tmn, and 5tln are included in the Workspace.
3. If the Calculate similarity panel is not open, open it by selecting Similarity from the Strike submenu of the Applications menu.
4. Select Descriptor similarities.

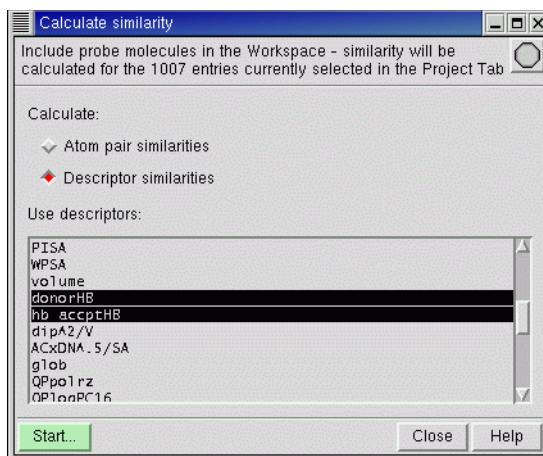


Figure 3.24. The Calculate Similarity panel with descriptor similarities specified

The Use descriptors list becomes available for choosing molecular properties to use in calculating similarities.

5. Select the donor HB and hb accpthB descriptors as shown in [Figure 3.24](#) and click Start.

The Statistics / Similarity - Start dialog box opens.

6. Change the job options if necessary, then click Start to run the calculation.

After a few seconds, the Monitor panel reports that the job has finished. In the Project Table, the calculated descriptor similarities have been added as properties: Euclidean Similarity, Euclidean Squared Similarity, Tanimoto Similarity, and Manhattan Similarity.

In the descriptor similarity calculation, the molecular properties of the probe molecules are averaged so they can be treated as a single virtual probe molecule. Similarity is calculated with respect to only the selected properties. You will now examine how well descriptor similarity based only on the number of hydrogen-bond acceptors and donors extracts known actives from the data set.

7. Click the Sort button on the Project Table panel toolbar.



The Sort Project Table panel opens.

8. Select Euclidean Squared Similarity, set the Order to Ascending, and click Sort All Rows.

The smallest values, corresponding to the greatest similarity to the probes, appear at the top of the table. Of the 9 active ligands, 8 are found within the top 375 ligands. The remaining active, 1tlp, is ranked last at 1007, due to its very large number of hydrogen-bond acceptor sites.

9. Now sort based on Tanimoto Similarity in Descending order.

Of the 9 actives, 8 are found in the top 269 compounds. The 1tlp ligand is again the lowest-ranked active; in this case it is found at 556.

10. Close the Sort Project Table and the Calculate Similarity panels.

These very simple examples were designed to show possible applications of Strike similarity calculations in descriptor and 2D-similarity space.

3.5 Estimating Activity by Creating a QSAR Model

In this tutorial, you will build a QSAR model and use it to predict activity. The most significant difference between this exercise and the QSPR model-building exercise in [Section 3.2 on page 30](#) is that the property being predicted is a biological activity. The workflow and steps that follow are similar to the QSPR exercise.

3.5.1 Thymidylate Synthase Activity of Folate-Based Inhibitors

Thymidylate synthase is an anticancer drug target as it catalyses the generation of deoxy-thymidine monophosphate given dUMP and a cofactor, 5,10-methylene tetrahydrofolate, an essential step in de novo DNA replication. The widely used anticancer agent 5-fluorouracil targets thymidylate synthase and is active against solid tumors like breast, head, neck, and colon cancers. Activities (EC_{50} and IC_{50}) have been experimentally determined for a large series of compounds. You will use the set of broadly applicable descriptors generated by QikProp in order to develop a QSAR model with Strike.

For this tutorial, a set of 188 known inhibitors were selected. All of these ligands have experimentally-determined L1210 IC_{50} activities that range from 141 to 0.00052 μ M. This set of ligands is well suited for QSAR as the structures have similar cores with a large variety of substitution in shared sidechains which leads to a wide activity range. The ligands were prepared from 2-D geometries using LigPrep, then neutralized, prior to being run through QikProp to produce 36 predicted properties.

The Maestro format file `thymidylate_synthase_ligands.mae` contains the 188 ligand structures with their QikProp properties, their raw IC_{50} s, and their $-\log(IC_{50})$ s. Because the goal is a free energy relationship between activities and properties, you will use the $-\log(IC_{50})$ or $\log(1/IC_{50})$ values rather than the raw IC_{50} values.

3.5.2 Changing Maestro Directories

If you are starting a new Maestro session:

1. Change to your *working-directory/qsar/* directory
2. Start Maestro by entering

```
$SCHRODINGER/maestro &
```

The Maestro main window is displayed.

If you are already in a Maestro session:

1. If there is an open project (not part of this tutorial), choose Close from the Project menu.
2. In the Display menu, ensure that Command Input Area is selected.

The Maestro main window includes the command input area.

3. In the Commands text box, enter the appropriate `cd` command to change to the directory *working-directory/qsar*. For example, if you are in *working-directory/simil*, enter the command:

```
cd ../qsar
```

3.5.3 Preparing the Data

The ligand data must be imported into the project and divided into a test set and a training set.

1. Click the Import structures button on the toolbar.



2. In the Import panel, select the file `thymidylate_synthase_ligands.mae`.
3. Ensure that Import all structures is selected, and that the Include in Workspace option selected is First Imported Structure.
4. Click Import.

After a moment, the first ligand in the file appears in the Workspace.

5. Close the Import panel, and open the Project Table panel.

There are 188 entries in the project, all selected.

6. Choose Random from the Select menu.

The Random Selection dialog box opens.

7. Ensure that the value in the Randomly select n % of entries text box is 50, the default, and click Select.

After a moment, the Project Table is redisplayed with half the entries deselected at random. The selected entries counter in the upper right corner of the panel now reads 94 selected.

To keep track of the newly selected entries, which will be used as the training set, add a column to the Project Table that labels the currently selected molecules:

1. Select Add from the Property menu to open the Add Property panel.
2. In the Name text box, type Population.
3. Choose String from the Type option menu.
4. In the Initial value text box, type training.
5. Click Add.

Under the column header Population, only the currently selected entries have a value of training.

3.5.4 Model Generation

You will now build a QSAR model employing all of the relevant QikProp descriptors:

1. Choose Build QSAR Model from the Strike submenu of the Applications menu.

The Build QSAR Model panel opens. The input counter under the title bar reads Input is 94 entries currently selected in the Project Table.

2. Under Select descriptors to be included in the model, select all the descriptors.
3. Control-click to deselect the following descriptors: Activity ($-\log[\text{IC}_{50}]$), #stars, #rtvFG, CNS, QPlogBB, and #metab.

The latter five descriptors are omitted because they are expected to be unrelated to the binding process; the first will be the dependent variable.

4. Ensure that the Regression method is Partial Least Squares and that Automatically remove outliers is not selected. See [Figure 3.25](#) to check your settings.
5. In the Maximum number of factors box, type 20.
6. Click Choose.

The Choose Activity Property dialog box opens.

7. Select Activity ($-\log[\text{IC}_{50}]$) from the list and click OK.

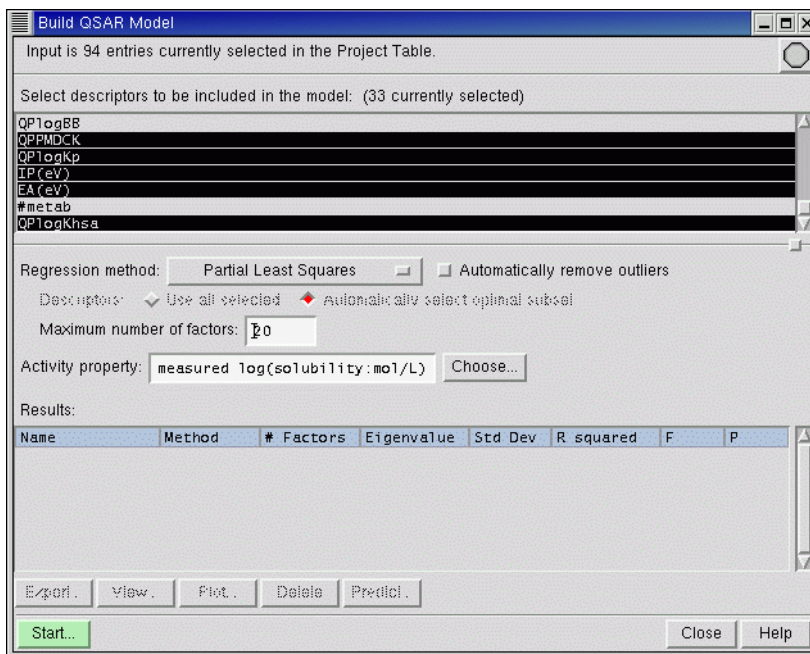


Figure 3.25. Build QSAR Model *panel showing settings for activity model*

- Click Start.

The Statistics / Build QSAR - Start panel opens.

- Change the job options if necessary, then click Start to begin the job.

The Strike job takes a few seconds to run.

When the job is finished, 20 potential models representing the 20 factors extracted are shown in the Results section of the Build QSAR Model panel. The predicted activities for all 20 factors are added to the Project Table under the headers Predicted ActivityX.Y.

With 20 factors, the model fit to the 94 molecules in the training set should have a high R squared, a large F-statistic and a very small P-factor. The standard deviation (Std Dev) should decrease from 1 to about 10 factors and then become somewhat constant while the R squared value should increase continuously as more factors are included, leveling off at about 10 factors. Thus, much of the predictive information is contained in the first 10 factors.

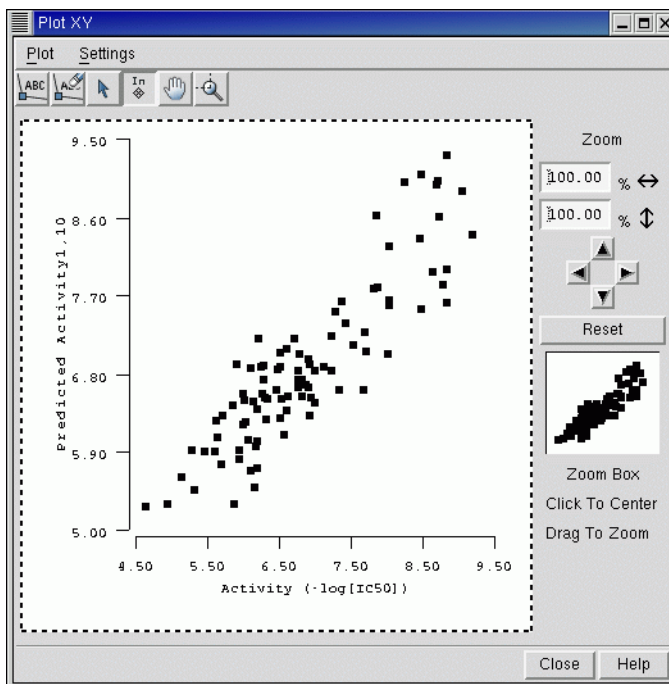


Figure 3.26. Plot of predicted vs. measured activity for training set

Inspect the results for the training set using the 10-factor predictor:

1. Select the 10-factor predictor (#Factors equal to 10) in the Results table.
2. Generate a plot of predicted activities versus experimental activities for the training set by clicking Plot. See [Figure 3.26](#).

3.5.5 Applying the Model to the Test Set

The true test of any model is to check its predictions against a set of molecules not included during its training. The exercise performed in this section would typically be considered part of model generation and validation, but for the purposes of this tutorial, it will be used to demonstrate the model application step of the Strike workflow.

The first step is to create a test set of molecules. In this example, the test set will be those molecules in the Project Table that were not members of the training set.

1. In the Project Table, confirm that the training set is selected by examining the Population column.

If so, skip to the next step. If for any reason the training set is no longer the selected set — for example, if a single entry has been selected instead — you can restore the selection by performing these steps:

- a. Choose Only from the Select menu of the Project Table.

The Entry Selection panel opens.

- b. In the Properties list, select Population.
- c. Select the option Is defined (any value).

Only the training set has a defined value (training) in the Population column.

- d. Click the Add button and then the OK button.

The molecules in the training set, and only those molecules, are now selected. The molecules that were in the training set cannot be part of the test set, so you will invert the selection.

2. In the Project Table, choose Invert from the Select menu.
3. In the Build QSAR Model panel, with the 10-factor predictor 1.10 selected, click Predict.

The Predict based on QSAR model panel opens showing the model with 20 predictors.

4. Ensure that the 10-factor predictor is selected as shown in [Figure 3.27](#) and click the Start button.
5. In the Statistics / Predict - Start dialog box, click Start to begin the calculation.
6. When the prediction job is finished, open the Project Table to view the new series of predicted activities for the test set.

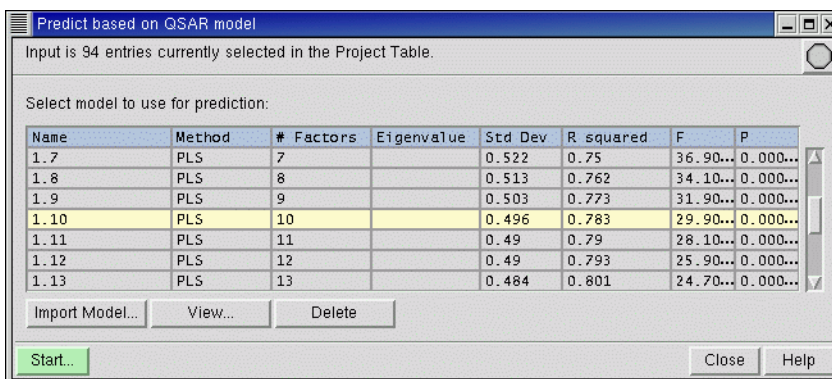


Figure 3.27. Predict based on QSAR Model *panel with 10-factor predictor selected*

Plot the results for the test set:

1. Open the Plot XY panel by clicking the Plot button in the Project Table toolbar



or selecting Plot from the Table menu.

The Plot XY panel opens with the most recent plot (generated in [Section 3.5.4 on page 62](#)) displayed.

2. Choose Delete Selected Plots from the Plot menu.

The plot is deleted.

3. Choose New Plot from the Plot menu.

The New Plot dialog box opens (see [Figure 3.10 on page 43](#)).

4. Change the following settings:

- a. In the Plot name text box, type test-activity.
- b. In the Series name text box, type tutorial.
- c. In the Axes section, select Activity (-log[IC50]) from the X property list.
- d. Select Predicted Activity1.10.10 from the Y property list.
- e. In the Drawing Style section, select None from the Line style list.

5. Click New to generate the plot.

By default, the Plot XY facility automatically sets the range for the X and Y axes of the plot. However, it is easier to spot outlying points when the X and Y axes share a common scale. If instead one spans a broader range of values than the other, this can be adjusted in the Plot Settings panel, as follows:

- a. Choose Plot Settings from the Settings menu to open the panel.
- b. Click the tab (X Axis or Y Axis) for the axis with the smaller range.
- c. In the X Axis or Y Axis folder, click the Edit button to open the Edit Axis dialog box.
- d. If Auto range is selected, deselect it to make the Minimum and Maximum text boxes available.
- e. Adjust the values to match the axis with the larger range, and click Edit to close the dialog box.

A sample plot is shown in [Figure 3.28](#).

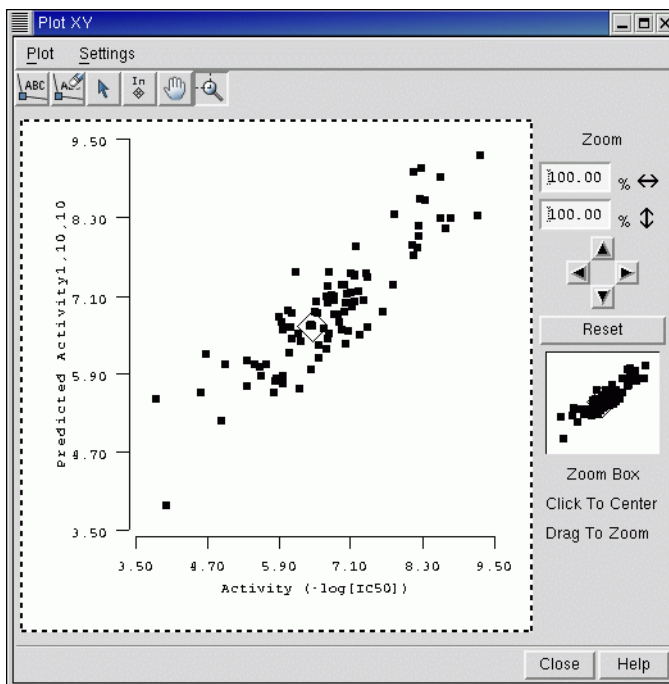


Figure 3.28. Plot of predicted vs. measured activity for the test set

The agreement between calculated and experimental values remains good for the test set. Individual data points of interest (outliers, high or low activity molecules) can be viewed in the Workspace by clicking the **In** button, then choosing data points.

This workflow could be repeated, using a different regression method to build the model (PCA or MLS), varying the independent descriptors in number or kind, or choosing different molecules for the training set; the QSPR exercises in [Section 3.2](#) give some brief examples.

This completes the QSAR model-building and application tutorial.

Running Strike from Maestro

Before using Strike, molecular data (also referred to as “descriptor data”) should be obtained and imported into the Maestro Project Table. This data can be generated using QikProp or other Schrödinger programs. Descriptors for ligands that bind to a receptor can be generated using the Ligand & Structure-Based Descriptors panel. This panel provides an interface to Liaison, Prime, MacroModel (eMBrAcE and `ligparse`), and QikProp to generate descriptors. For more information, see the document *Ligand and Structure-Based Descriptors*. Descriptors generated by external programs or sources may be imported using standard comma-separated value (CSV) format files.

Once the data is incorporated in the Project Table, you can perform statistical analyses and create and use QSAR models using the Strike panels in Maestro.

The Strike interface in Maestro consists of three panels:

- **Build QSAR Model**—Generate a QSAR model using a training set of molecules selected from the Maestro Project Table, a set of independent descriptors, and a dependent descriptor chosen from those available in the data.
- **Predict based on QSAR model**—Import a model or select one from the table of generated models, then perform property predictions for molecules that were not part of the training set. Results can be viewed and unsatisfactory models can be deleted from the table.
- **Similarity**—Determine similarities in descriptor or 2D-structure space. Several distance-based descriptor similarity measures are available. Similarity in 2D-structure space is determined using an atom-pair-based approach.

4.1 The Build QSAR Model Panel

To open the Build QSAR Model panel, choose Build QSAR Model from the Strike submenu of the Applications menu. It may also be useful to open the Project Table containing your molecular data.

4.1.1 Using the Build QSAR Model Panel

Use the panel to generate a QSAR model given a training set of molecules selected from the Project Table, a set of independent descriptors, and a dependent variable for which a prediction will be made. Three regression techniques are available: multiple linear regression (MLR),

partial least squares regression (PLS), and principal component analysis (PCA). Certain options are available only for the appropriate regression method.

When you have finished selecting options, click the green Start button to open the Start dialog box for the Build QSAR Model job. Choose the Host machine and the appropriate Username, then click Start.

As the job starts, the gray octagon in the upper right corner of the panel turns green and begins to rotate, indicating that a job is in progress, and the job Monitor panel is displayed. The running log for the job appears in the Monitor panel. If you have closed the Monitor panel and want to reopen it, click the octagon (or choose Monitor Jobs from the Applications menu.) Most model-building jobs take a few seconds to complete.

Once a model has been built, 2D plots of predicted properties versus the dependent descriptor data can be created. Clicking points in a plot brings the molecule or molecules selected into the Maestro 3D Workspace for viewing and manipulation. (For more information about these and other Maestro plots, see the Plot XY Panel topic in the Online Help.)

From the Build QSAR Model panel you can proceed directly to the Predict based on QSAR model panel or save the generated models in the project for later use.

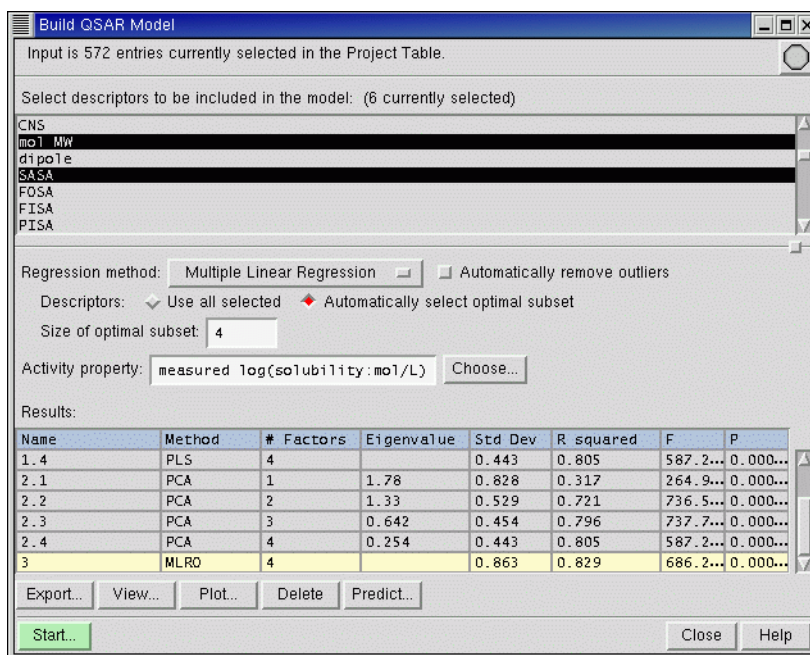


Figure 4.1. The Build QSAR Model panel

4.1.2 Build QSAR Model Panel Features

- Input is N entries currently selected in the Project Table

The entries that are selected (highlighted) in the Project Table will be used to build the QSAR model. The number of entries selected is displayed in the upper portion of the panel.

Optionally, you can use the Random option in the Select menu of the Project Table to randomly select a specified percentage of either the selected or the total entries. The default is 50% of the selected entries.

- Select descriptors to be included in the model list

Choose an appropriate set of independent descriptors that is likely to correlate with the dependent descriptor and a regression method by selecting them in the list.

- Regression Method option menu

The options for regression method are:

- Partial Least Squares (PLS)

When this method is selected, the Maximum number of factors option becomes available. The range for Maximum number of factors is from 1 to the number of selected descriptors. The number of molecules selected to be used in building the model must be greater than or equal to the maximum number of factors. For more information about the method, see [Section 6.3.2 on page 91](#).

- Principal Component Analysis (PCA)

When this method is selected, the Maximum number of factors option becomes available. The range for Maximum number of factors is from 1 to the number of selected descriptors. The number of molecules selected must be greater than or equal to the number of descriptors chosen. For more information about the method, see [Section 6.3.3 on page 91](#).

- Multiple Linear Regression (MLR)

When this method is selected, the Descriptors options become available. The number of molecules selected must be greater than or equal to the number of initial descriptors. It is recommended that the number of molecules be at least five times greater than the number of descriptors. For more information about the method, see [Section 6.3.4 on page 92](#).

- Automatically remove outliers

Select this option to remove outlying molecules before the model is built. By default, this option is not selected and outliers are not removed. For samples of 500 members or more, selecting this option will greatly increase the time needed for the model-building job. See [Section 6.5 on page 93](#).

- Descriptors options (MLR)

When the regression method selected is MLR, you can choose to Use all selected descriptors (command-line keyword value MLRS) or to Automatically select optimal subset of selected descriptors (command-line keyword value MLRO).

- Maximum number of factors text box (PLS, PCA)

When the regression method selected is PLS or PCA, this option is available.

- Size of optimal subset text box (MLR)

When the regression method selected is MLR, this option is available.

- Activity Property text box and Choose button

Click Choose to open a list of all the descriptors in the data, from which you can select the property you want the model to predict.

- Results table

This table lists the models which have been calculated in the current Maestro project. Select a single model to export, view, plot, delete, or use for prediction. Along with the name of each model, the table includes the regression method used, the number of PLS/PCA factors or MLR descriptors, the eigenvalue for PCA models, and standard statistics values.

- Export

Export the currently selected model to an external file.

- View

Click View to review the output file of the model-building job for the selected model, which contains all the data needed to completely describe the model.

- Plot

Generate a Maestro Plot XY plot of the predicted versus experimental activity for the currently selected model.

- Delete

Delete the currently selected model from the table.

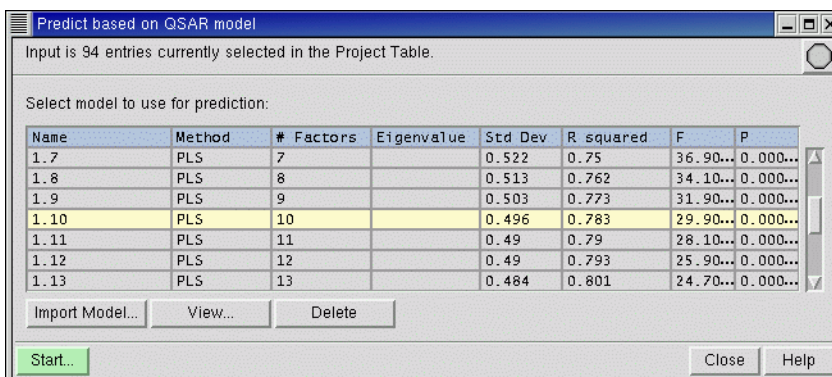


Figure 4.2. The Predict based on QSAR Model *panel*.

- Predict

Open the Predict based on QSAR model panel with the currently selected model chosen.

4.2 The Predict Based on QSAR Model Panel

To open the Strike Predict based on QSAR model panel, choose Predict from the Strike submenu of the Applications menu. The Predict based on QSAR model panel can also be opened from the Build QSAR Model panel once one or more models have been generated, using the Predict button in the lower portion of the panel.

4.2.1 Using the Predict Panel

Models to be used for property predictions can be imported from another project or generated in the current project. You must have data for all independent descriptors used in the model. To make predictions, select the desired molecules in the Project Table. Pick the model to use and click Start to generate predictions which are automatically imported into the Project Table.

4.2.2 Predict Panel Features

- Input is N entries currently selected in the Project Table

The selected model will be used to predict properties of the entries that are selected (highlighted) in the Project Table.

- Select model to use for prediction table

This table contains information about each model in the current project. Choose a single model for the prediction job. When the job is complete, the table will display the predicted property values.

- Import Model button

Import a model into the table. For example, you can import a model that was exported from the Build QSAR Model panel in an earlier Strike session.

- View button

Click View to review the output file for the selected model, which contains all the data needed to completely describe the model.

- Delete button

Delete the currently selected model from the table.

4.3 The Calculate Similarity Panel

To open the Calculate similarity panel, choose Similarity from the Strike submenu of the Applications menu.

4.3.1 Using the Calculate Similarity Panel

Use this panel to determine similarities in descriptor or 2D-structure space. Several distance-based descriptor similarity measures are available. Similarity in 2D-structure space is determined using an atom-pair-based approach.

Similarity, either atom-pair-based or descriptor-based, is calculated with respect to probe molecules. Select probe molecules by including them in the Workspace. At least one molecule must be included in the Workspace, and at least one entry must be selected in the Project Table, before similarity calculations can proceed.

4.3.2 Calculate Similarity Panel Features

- Similarity will be calculated for the N selected entries

The number of entries selected in the Project Table is displayed. Similarity will be calculated for the selected entries only.

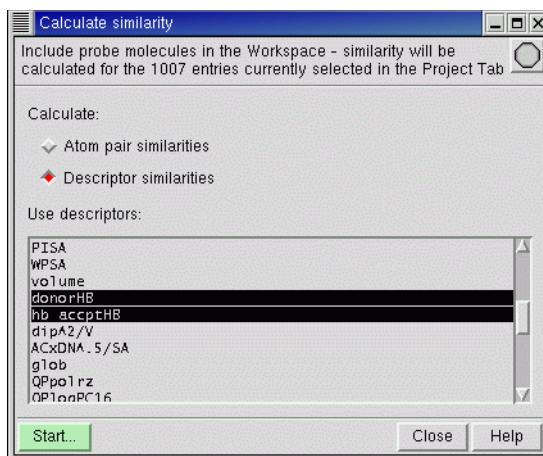


Figure 4.3. The Calculate Similarity panel with descriptor similarities specified.

- Calculate options

The options for similarity calculations are:

- Atom pair similarities

A similarity property will be created for each selected entry in the Project Table.

- Descriptor similarities

Similarity in terms of properties will be calculated for the selected Project Table entries in terms of the properties chosen from the Use Descriptors scrolling list.

- Use Descriptors table

This list is available when descriptor similarities are selected to be calculated. Choose the properties you want to include in the similarity calculation.

Running Strike from the Command Line

Strike can also be run using the `strike` command and input files in `.csv` format. This chapter lists keywords and gives two examples of command blocks that can be used in input files.

5.1 Usage Summary

```
$SCHRODINGER/strike [options] inputfile
```

<i>inputfile</i>	The <code>strike</code> input script file containing the commands to be performed.
<code>-HOST host</code>	Run job on a remote host.
<code>-LOCAL</code>	Run the job in the current directory, rather than in a temporary scratch directory.
<code>-WAIT</code>	Do not return until the job completes.
<code>-NICE</code>	Run the job at reduced priority.
<code>-HELP</code>	Print this message and exit.

5.2 Input File Examples

The `strike` input script consists of blocks of commands, each consisting of a series of `keyword=value` pairs and terminated by a line beginning with `#`. The termination line beginning with `#` is mandatory, even if there is only one block of data in the input script. Each command block is then executed sequentially. Comment lines must begin with `!!`. Two examples of command blocks are given below.

```
!! Section to train a model
dataFile=input_files/strike_mlro_fit.csv
runMode=train
model=MLRO
activityLabel=activity
numOptDescript = 4
# End of Section 1

!! Section to use a previously created model
runMode=test
dataFile=input_files/strike_pls_fit.csv
modelFile=input_files/strike_pls_fit.model
# End of Section 2
```

5.3 Input File Keywords, Values, Descriptions

The following *keyword=value* pairs are accepted input for `strike`. Boolean values must be expressed as `yes` or `no`.

5.3.1 Mode Selection

All Strike jobs must use the `runMode` keyword with one of these values.

Value	Description
<code>train</code>	Generate a QSAR model.
<code>test</code>	Predict properties using a QSAR model.
<code>simil</code>	Run a descriptor similarity calculation.
<code>apsimil</code>	Run an atom-pair (2D structure space) similarity calculation.
<code>stats</code>	Generate statistics.
<code>factorGen</code>	Extract factors from PCA model.
<code>factorRed</code>	Reduce data using extracted PCA factors.
<code>factorExp</code>	Expand reduced data using extracted PCA factors.

5.3.2 File Specification Commands

Keyword	Value	Description/Relevant Job Types
<code>dataFile</code>	<code>datafilename</code>	Keyword required for all jobs except atom-pair similarity (<code>apsimil</code>). File must be in CSV format.
<code>outputFile</code>	<code>outputfilename</code>	Default is <code>jobname.out</code> . All jobs.
<code>modelFile</code>	<code>modelfilename</code>	Default is <code>jobname.model</code> . File containing QSAR model. QSAR jobs (<code>train</code> and <code>test</code>) and factor reduction jobs (<code>factorGen</code> , <code>factorRed</code> , <code>factorExp</code>).
<code>csvFile</code>	<code>csvoutfilename</code>	Default is <code>jobname.csv</code> . Output file containing all data used and generated in current command block. All jobs except <code>apsimil</code> .
<code>plotFile</code>	<code>qsaroutfilename</code>	File containing output from <code>train</code> with predicted vs. dependent data. QSAR jobs.
<code>apPredFile</code>	<code>filename</code>	File of molecules whose similarity to the probes are to be determined. <code>apsimil</code> jobs.
<code>apActivesFile</code>	<code>activesfilename</code>	File of probe molecules. <code>apsimil</code> jobs.
<code>apInactivesFile</code>	<code>inactivesfilename</code>	File of decoy molecules. <code>apsimil</code> jobs.
<code>apWeightsFile</code>	<code>weightsfilename</code>	Weights file. When generated, default is <code>jobname.csv</code> . <code>apsimil</code> jobs.

5.3.3 Alternative Naming Convention Commands

Keyword	Value	Description/Relevant Job Types
modelTitle	<i>modelname</i>	Alternative title for QSAR model generation. Otherwise defaults to <i>job-name</i> .
baseName	<i>basename</i>	Alternative basename for all jobs. All output files will be <i>basename</i> .*, and modelTitle will default to <i>basename</i> .

5.3.4 Commands for Reading/Writing .csv Files

Keyword	Value	Description/Relevant Job Types
delim	<i>string</i>	Delimiter character for reading .csv file. All jobs except apsimil.
includeColumns	<i>X:Y, Z</i> column numbers or column labels	X, Y, Z can be numbers or labels (headers). Use colon for ranges (e.g., includeColumns=2:6,9,15 includes columns 2-6 and 9 and 15 from the input file). All jobs except apsimil.
excludeColumns	<i>X:Y, Z</i> column numbers or column labels	X, Y, Z can be numbers or labels (e.g., labels: excludeColumns=IP(ev):QPlogKhsa Properties in .csv file between IP(ev) and QPlogKhsa, inclusive, will not be used). All jobs except apsimil.
includeRows	<i>X:Y, Z</i> row numbers or row labels	Include molecules (rows) specified. All jobs except apsimil.
excludeRows	<i>X:Y, Z</i> row numbers or row labels	Exclude molecules (rows) specified (e.g., excludeRows=25 excludes molecule 25). All jobs except apsimil.
activityColumn	<i>integer</i>	Identify dependent property by column number. Build QSAR model jobs.
activityLabel	<i>Label</i>	Identify dependent property by column label. Build QSAR model jobs.
rowHeaderColumn	<i>integer</i>	Set the column in the .csv file that contains row labels, by column numbering beginning at 1. For all jobs if needed.

Keyword	Value	Description/Relevant Job Types
rowHeaderLabel	<i>Label</i>	Set the column in the .csv file that contains row labels by column label. For all jobs if needed.
descriptorWeightRow	<i>integer</i>	Set by number the row that contains the weight for each descriptor. For descriptor similarity jobs.
descriptorWeightLabel	<i>Label</i>	Set by label the row that contains the weight for each descriptor. For descriptor similarity jobs.

5.3.5 Commands for Build QSAR Model (train) Jobs

Keyword	Value	Description/Relevant Job Types
model	PLS PCA MLRS MLRO NNET	Specify type of regression to be employed.
autoScale	yes no	Set whether data is to be converted to a common scale. Default is yes.
maxFactors	<i>integer</i>	Maximum number of factors to return. PLS, PCA.
numOptDescript	<i>integer</i>	Number of descriptors to be retained, determined by optimization. MLRO.
removeOutliers	no yes	Run prior to importing data into model building. Compare relative densities in descriptor space for included molecules to predict outliers. Recommended for sample size < 500 only. Default is no.
printMLROutliers	no yes	Set to yes to output possible outliers with respect to the MLR model. Default is no. MLRS, MLRO.
MLROutlierCutoff	<i>integer</i>	Integer from 0 to 5 giving the number of MLR outlier tests that need to fail before a data point is identified as a possible model outlier. Default is 4. MLRS, MLRO.
lgoPercent	double	For leave-group-out (LGO) validation, percentage of fitting set to use as test set for each regression. Default is 5.0%
lgoCycles	<i>integer</i>	Number of cycles of LGO validation to perform. Default is 10.
RandCycles	<i>integer</i>	Number of randomization cycles to perform. Default is 10 times the number of independent descriptors. MLRS, MLRO, PLS, and PCA.
supYintercept	no yes	Suppress inclusion of the y intercept as a dependent variable for regression generation. The default is no, which means that the y intercept is included. MLRS, MLRO.

Keyword	Value	Description/Relevant Job Types
nnetNumUnitsInHidden Layer	<i>integer</i>	Number of units in the hidden layer. NNET.
nnetCrossValPer	<i>integer</i>	Percent of input data to be kept in the cross validation set. Default is 5%. NNET.
nnetExtValPer	<i>integer</i>	Percent of input data to be kept in the external validation set. Default is 10%. NNET.
nnetNumTrainCycles	<i>integer</i>	Number of training cycles for each neural network. Default is 200. NNET.
nnetNumNetworks	<i>integer</i>	Number of neural networks to train of which the best nnetumNetworksEnsem will be selected to create an ensemble neural network that is presented to the user. Default is 20. NNET.
nnetNumNetworksEnsem	<i>integer</i>	Number of the best neural networks to use in generating an ensemble neural network that is presented to the user. Default is 5. NNET.

5.3.6 Commands for Atom-Pair Similarity (apsimil) Jobs

Keyword	Value	Description
apPredFormat	mae sdf	Format of apPredFile.
apActivesFormat	mae sdf	Format of apActivesFile.
apInactivesFormat	mae sdf	Format of apInactivesFile.
probes	<i>X:Y, Z</i>	Specify probe molecules. X, Y, Z can be molecule numbers or, if rowHeaderColumn is defined, molecule titles.
inactivePercent	<i>nn.n</i>	Percentage of inactives, e.g., for 99%: inactivePercent=99.0
readWeights	yes no	Read weights from apWeightsFile.
genWeights	yes no	Generate weights in apWeightsFile.
normalize	range z-score none	Specify normalization approach. Default (range) normalizes data to 0.0 - 1.0 scale; required for Tanimoto coefficient calculation. Specify z-score to scale data in standard deviation units. Specify none to perform no normalization.

5.3.7 Commands for Factor Reduction Jobs

Keyword	Value	Description
facRedAuto	yes no	Determines if the factors are to be generated using scaled (yes) or unscaled (no) input data.
facRedNumFactors	<i>n</i>	Number of factors to generate. Range is from 0 to the number of input data columns.

5.3.8 Other Commands

Keyword	Value	Description/Relevant Job Types
enrich	Euclidean Euclidean_sq Tanimoto Manhattan	Calculate enrichment factors for extracting probe molecules from the entire data set, using the specified similarity measure. For <i>simil</i> and <i>apsimil</i> jobs.
stats	<i>label</i>	Calculate univariate statistics for the descriptor <i>label</i> . Any job.
	<i>label1</i> , <i>label2</i>	Calculate bivariate statistics for the descriptors <i>label1</i> and <i>label2</i> . Any job.

5.3.9 Keyword Requirements for Various Job Types

Table 5.1. Minimum Strike Keywords by Job Type

Job Type	Keywords Required	Comments
Build QSAR model	Keywords depend on chosen model.	Column containing dependent data can be specified by column number (activityColumn) <i>or</i> by column label (activityLabel).
Build QSAR model Partial Least Squares	runMode=train model=PLS dataFile activityColumn <i>or</i> activityLabel maxFactors	
Build QSAR model Principal Component Analysis	runMode=train model=PCA dataFile activityColumn <i>or</i> activityLabel maxFactors	

Table 5.1. Minimum Strike Keywords by Job Type (Continued)

Job Type	Keywords Required	Comments
Build QSAR model Multiple Linear Regression Analysis	runMode=train model=MLRS dataFile activityColumn <i>or</i> activityLabel	
Build QSAR model MLRS with optimum number of descriptors	runMode=train model=MLRO dataFile activityColumn <i>or</i> activityLabel numOptDescript	
Build QSAR model Neural Network	runMode=train model=NNET dataFile activityColumn <i>or</i> activityLabel nnetNumUnitsInHidden Layer	
Validation or predic- tion using any model	runMode=test dataFile modelFile	Model is entirely specified in modelFile.
Descriptor similarity calculation	runMode=simil dataFile probes (rowHeaderColumn <i>or</i> rowHeaderLabel)	One of the keywords in parentheses needed to specify probe molecules if probes not specified by molecule number.
Atom-pair similarity calculation. No weights	runMode=apsimil apActivesFile apActivesFormat apPredFile apPredFormat	
Atom-pair similarity. Weights are generated and used in same command block	runMode=apsimil apActivesFile apActivesFormat apPredFile apPredFormat apInactivesFile apInactivesFormat inactivePercent genWeights	

Table 5.1. Minimum Strike Keywords by Job Type (Continued)

Job Type	Keywords Required	Comments
Atom-pair similarity using weights that were generated previously	runMode=apsimil apActivesFile apActivesFormat apPredFile apPredFormat readWeights apWeightsFile	
PCA factor generation	model=PCA runMode=factorGen facRedNumFactors facRedAuto=yes dataFile	Also generates reduced data set for the input.
PCA factor reduction	model=PCA runMode=factorRed facRedNumFactors facRedAuto=yes dataFile modelFile	modelFile must be output of a factor generation job. dataFile must contain the same descriptors as used in the factor generation, but need not contain data for the same structures.
PCA factor expansion	model=PCA runMode=factorExp facRedNumFactors facRedAuto=yes dataFile modelFile	modelFile must be output of a factor generation job. dataFile must be a file with reduced factors from a factor generation or reduction job.

Statistical Definitions and Methods

This chapter defines statistical quantities, algorithms, and regression methods used in Strike.

6.1 Univariate Statistics

This section defines some symbols, definitions, and equations relating to the statistics of a single variable.

6.1.1 Symbols

N

Number of data points (observations) in a sample. There is no hard limit on sample size (number of molecules) in Strike, but for large samples (millions of molecules) practical issues such as system memory limitations may apply.

x_i —value of data point i in a sample of variable x

\bar{x} —mean of variable x

6.1.2 Mean, Median, and Mode

These statistics describe the “central tendency” of a variable.

Mean

The *mean* of variable x is defined by [Equation \(1\)](#).

$$\bar{x} = \sum_i^N x_i / N \quad (1)$$

Median

- For an even number of data points, the *median* is the mean of the middle-most two values in the ordered sample.
- For an odd number of data points, the *median* is the middle-most value in the ordered sample.

One-half of the ranked values for a variable will lie above and one-half below the value of the median.

Mode

The *mode* is the value of a variable that occurs with the greatest frequency in a sample. If more than one value shares the highest frequency of occurrence, the term “mode” is not applicable.

6.1.3 Variance and Deviation

The statistical quantities defined in this section are measures of the spread of values in a sample about the mean.

Variance

The *variance* is defined as:

$$\sigma^2 = \sum_i^N (x_i - \bar{x}) / (N - 1) \quad (2)$$

If \bar{x} is known or if one is examining a complete population, the $N-1$ term reverts to N . Strike calculates all variances assuming a sample, i.e. using $N-1$, which is suitable for the vast majority of cases. The *mean squared deviation*, also defined in this section, reports the variance for a population.

Standard Deviation

The *standard deviation* is the square root of the variance, defined in [Equation \(2\)](#).

If \bar{x} is known or if one is examining a complete population, the $N-1$ term reverts to N . Strike calculates all standard deviations assuming a sample, i.e. using $N-1$, which is suitable for the vast majority of cases. The *root mean squared deviation*, also defined in this section, reports the standard deviation for a population.

The standard deviation measures the spread of values about the mean. If the sample exhibits a normal distribution, then 68.3% of values will lie within 1σ from the mean, 95.4% of values will lie within 2σ of the mean, and 99.7% of values will lie within 3σ of the mean.

Mean Absolute Deviation

$$\text{MAD} = \sum_i^N |x_i - \bar{x}| / (N - 1) \quad (3)$$

Mean Squared Deviation

$$\text{MSD} = \sum_i^N (x_i - \bar{x})^2 / N \quad (4)$$

If \bar{x} is known or if one is examining a complete population, the *mean squared deviation* reports the variance for the complete population.

Root Mean Squared Deviation

$$\text{RMSD} = \sqrt{\sum_i^N (x_i - \bar{x})^2 / N} \quad (5)$$

If \bar{x} is known or if one is examining a complete population, the *root mean squared deviation* reports the standard deviation for the complete population.

6.1.4 Skewness and Kurtosis

These statistics are measures of the extent to which a sample differs from a normal distribution. Both have a value of zero for a normal distribution.

Skewness

$$\mu = \frac{\sum_i^N (x_i - \bar{x})^3 / N}{\text{RMSD}^3} \quad (6)$$

Strike calculates the Fisher Skewness for a sample. Normal distributions have a skewness of zero as they are perfectly symmetrical about the mean. A positive value of the skewness indicates, relative to a normal distribution, that the sample being examined is asymmetric and skews towards larger values, i.e. has a larger tail to the right. A negative value of the skewness indicates, relative to a normal distribution, that the sample being examined skews toward smaller values, i.e. has a larger tail to the left. A significant skewness value indicates that the sample does not have a normal distribution.

Kurtosis

$$\text{kurtosis} = \frac{\sum_i^N (x_i - \bar{x})^4 / N}{\text{RMSD}^4} - 3 \quad (7)$$

Strike calculates the excess kurtosis using the formula of Snedecor and Cochran. The kurtosis for a normal distribution is three. By subtracting three, Strike reports the excess kurtosis where a normal distribution has a kurtosis of zero. A positive kurtosis indicates the distribution is strongly peaked about the mean while a negative kurtosis indicates the distribution is flat. A significant kurtosis value indicates the sample does not have a normal distribution.

6.2 Bivariate Statistics: Covariance and Correlation

The statistics in this section describe the relationship between two variables in terms of covariance and correlation. These statistics are also applied to pairs of variables in models with multiple independent variables.

Correlation coefficient

See *Pearson r* or *r-squared*

Correlation matrix

The *correlation matrix* generates the Pearson *r* values for the half matrix of all pairs of selected variables.

Covariance

$$\text{cov}(x, y) = \frac{\sum_i^N (x_i - \bar{x})(y_i - \bar{y})}{N - 1} \quad (8)$$

The covariance measures the extent to which two variables vary together. A positive value of the covariance indicates that larger than average values of one variable tend to be paired with larger than average values of the second variable. A negative value of the covariance indicates that larger than average values of one variable tend to be paired with smaller than average values of the second variable. A zero covariance indicates the two variables vary independently from one another. The covariance is dependent on the magnitude of the variables involved and is most useful when the variables have the same magnitude.

For a scatter plot of x and y the covariance measures how close the scatter is to a line. A negative covariance indicates a downward sloping line to the right, a positive covariance indicating an upward sloping line to the right, and a zero covariance indicating the best line lies along the horizontal axis.

Pearson r

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} \quad (9)$$

The *Pearson r* is a correlation coefficient that determines the extent that two variables are proportional to one another. In other words, the Pearson r provides a measure of linear association between variables. Calculated Pearson r values lie on a scale from -1.0 to +1.0 with negative values indicating the best least-squares line between variables x and y is downward sloping to the right and positive values indicating the best line is upward sloping to the right. A value of zero indicates no correlation between the two variables. The Pearson r is independent of the magnitude of variables (unlike the covariance). Note that R is sometimes used instead of r .

R-squared

In *Strike*, *r-squared* is the square of the Pearson r correlation coefficient. Its value ranges from 0.0 to 1.0 with a value of zero indicating the two variables have no correlation and a value of one indicating the variables are perfectly correlated. Like the Pearson r , the *r-squared* is independent of the magnitude of the two variables.

Spearman rho

The *Spearman rho* is a rank-order correlation coefficient. It measures the proportion of variability accounted for between two variables using the ranking of the data rather than the data values themselves. The Spearman rho is interpreted in an identical fashion to the Pearson r statistic.

Ties in ranking (data points with the same value) are given the mean rank of the tied observations. i.e. if three points are identified as having equal values with ranks of 5, 6, 7, and 8 in the sample, the average rank assigned to all four would be 6.5. In the definitions below, $\text{rank}(x_i)$ is the rank of the point x_i , $\text{ties}(x_i)$ is the number of times the value x_i occurs, and in $\epsilon(x)$ the sum is over the number of tied values.

$$D = \sum_i^N [\text{rank}(x_i) - \text{rank}(y_i)]^2 \quad (9a)$$

$$\epsilon(x) = \sum_i \text{ties}(x_i)^3 - \text{ties}(x_i) \quad (9b)$$

$$\rho = \frac{1 - [6D + (\epsilon(x) + \epsilon(y))/2] / (N^3 - N)}{\sqrt{1 - \epsilon(x)/(N^3 - N)} \sqrt{1 - \epsilon(y)/(N^3 - N)}} \quad (9c)$$

Kendall tau

The *Kendall tau* is a rank-order correlation coefficient. It measures the proportion of variability accounted for between two variables using the ranking of the data rather than the data values themselves. The Kendall tau is interpreted in an identical fashion to the Pearson *r* statistic. It is defined in Equation (10), where:

P is the number of concordant pairs of ranks

Q is the number of discordant pairs of ranks

*Y*₀ is the number of ties in the ranks of two *x*'s

*X*₀ is the number of ties in the ranks of two *y*'s

$$\tau_b = \frac{P - Q}{\sqrt{P + Q + X_0} \sqrt{P + Q + Y_0}} \quad (10)$$

To calculate the Kendall tau the half matrix of data pairs is analyzed, i.e. (*x_p*, *y_i*) and (*x_p*, *y_j*) are compared for all *i* and *j* pairs. Each pair that shows the same rank order between the two data sets is counted as concordant. Each pair that shows a different rank order between the two data sets is counted as discordant. The rank order can be determined by the following expression:

$$\begin{aligned} (\text{rank}(x_i) - \text{rank}(x_j))(\text{rank}(y_i) - \text{rank}(y_j)) &> 0 && \text{concordant} \\ &< 0 && \text{discordant} \\ &= 0 && \text{tie in } x \text{ or } y \end{aligned} \quad (11)$$

Ties are counted in the *Y*₀ and *X*₀ variables.

6.3 Model-Building Methods

This section defines terms and methods used in building QSAR/QSPR models. It briefly introduces the three regression methods available for model-building in Strike: partial least squares, principal component analysis, and multiple linear regression.

6.3.1 Independent and Dependent Variables

Dependent variable

The *dependent variable* (or *response variable*) is the variable that is being fitted to in a regression model. It is referred to as dependent as it is assumed that its values are dependent on the values of independent variables that will be used to generate the predictive model. In Strike, this variable is also referred to as the dependent descriptor or the activity property.

Independent variables

The *independent variables* are the variables that are being used to fit a regression to a dependent variable in partial least squares, principal component analysis, or multiple linear regression. They are referred to as independent as their values are assumed not to depend on the values of the dependent variable. In Strike, the term *independent descriptors* is often used.

6.3.2 Partial Least Squares

The *partial least squares* (PLS) method generates linear equations that describe the relationship between a number of factors derived from a set of independent descriptors and a dependent descriptor. The PLS procedure works by extracting successive linear combinations of the factors (also called components or latent vectors), which explain independent and dependent variations. In particular, the method of partial least squares balances these objectives, seeking factors that explain both response variation and predictor variation.

Partial least squares is particularly valuable because it can be applied in cases where the number of independent descriptors is greater than the number of molecules.

Partial least squares is similar to *principal component analysis*, but the goals of the two methods in extracting factors differ. In PLS one is concerned with the variance in both the dependent and independent descriptors, while in PCA one is trying to explain the maximum variance possible in only the dependent descriptors.

While it is possible to use PLS to generate models to fit multiple dependent variables, Strike is limited to fitting a single dependent variable.

6.3.3 Principal Component Analysis

Principal component analysis (PCA) transforms a number of independent variables into a number of uncorrelated factors that explain the variance of the dependent variable. The first factor accounts for as much of the variability in the data as possible, and each succeeding factor accounts for as much of the remaining variability as possible. The eigenvalues of the covariance matrix from PCA indicate the portion of the total variance accounted for by each

factor, where the total variance is generally defined as equal to the number of independent variables.

Principal component analysis can be applied in cases where the number of independent descriptors is greater than the number of molecules.

Principal component analysis is similar to partial least squares, but it focuses on explaining the maximum variance possible in only the dependent descriptors, while PLS considers the variance in both the dependent and independent descriptors.

While it is possible to use PCA to generate models to fit multiple dependent variables, Strike is limited to fitting a single dependent variable.

6.3.4 Multiple Linear Regression

Multiple linear regression (MLR) generates linear equations that describe the relationship between a set of independent descriptors and a dependent descriptor. Strike may only be used to fit a single dependent descriptor. As used in Strike, MLR fits a straight line to the dependent descriptor using the following linear relationship:

$$P_j = \sum_i c_i \chi_{ij} + c_0 \quad (12)$$

In the above equation, P_j is the property or activity that is to be predicted for each molecule j , the c_i values are the regression coefficients, χ_{ij} is the i th independent property for molecule j , and c_0 is a constant. Values of the coefficients and c_0 are fitted to give P_j values that reproduce the dependent value for the j th molecule.

In general, when fitting data using MLR it is advisable to use a data set with at least five times as many molecules as there are independent descriptors.

6.4 Model Analysis and Validation

The statistics in this section are used to analyze and validate QSAR/QSPR models built using regression techniques.

Cross validation or leave- n -out validation

Cross validation tests how dependent a generated regression is on the samples used to generate the regression. In leave-group-out (LGO) or leave- n -out cross validation, the original set of samples is divided into n subsets. Then, n regressions are generated, each time omitting a different subset. Each of the n regressions is then used to predict the expected dependent value for the molecules in the omitted subset. In n regressions all molecules will have had their

dependent value predicted and the r-squared from comparing the predicted dependent values against the true dependent values is referred to as the *q-squared*. To reduce the dependence of cross validation on the composition of the subsets randomly generated, the cycle is repeated *c* times. The mean of the *c* values of q-squared is reported by Strike. A q-squared value that deviates significantly from the r-squared for a regression generally indicates that the regression is overly dependent on the set of molecules included in the training set and may not have the desired predictive power.

By default, Strike uses a subset size (`lgoPercent`) of 5% of the sample, giving *n* subsets = 20, and a number of cycles *c* (`lgoCycles`) = 10. When Strike is run from the command line, the `lgoPercent` and `lgoCycles` keywords can be used to specify non-default values. See [Section 5.3.5 on page 80](#).

F-statistic

The *F-statistic* is used in regression analysis to determine if the variances between the means of two populations are significantly different. In other word, the F-statistic provides an indication of the lack of fit of the data to the estimated values of the regression. A strong relationship between two variables gives a high F-ratio.

Leave-*n*-out validation

See *cross validation*.

P-value

The *p-value* is the probability that the regression was obtained not from correlations between the dependent and independent variables, but instead by chance. Generally p-values of < 0.05, which indicate a 1 in 20 probability that the regression was obtained by chance, are considered statistically significant.

Q-squared

The *q-squared* is the r-squared determined by comparing the dependent variable against predictions made using a model. See *cross validation* for details.

6.5 Outlier Detection

Local Correlation Integral Outlier (LOCI) Detection

The LOCI outlier detection methodology uses a density-based approach to identifying outliers within a sample. It works by comparing the density of points surrounding a given point with the densities of the surrounding neighbor points. Significant differences in densities lead to the identification of outliers. It provides an automatic, data dictated cut-off to identify outliers

without the need for user input. This method does not suffer from either the local density problem or the multi-granularity problem. It should be noted that our implementation of the LOCI algorithm does not scale well for larger sample sizes, and as such should only be used on samples of less than about 500 members.

6.6 Similarity Statistics

The statistics defined in this section are measures of similarity.

6.6.1 Atom-Pair Similarity

m_{AB} is the total number of unique atom pair types found on molecules A and B

freq_k^A is the number of times atom pair type k was found on molecule A

w_i is the weight for atom pair type k

$$\text{sim}_{AB} = \frac{\sum_k^{m_{AB}} w_k \min(\text{freq}_k^A, \text{freq}_k^B)}{0.5 \sum_k^{m_{AB}} w_k (\text{freq}_k^A + \text{freq}_k^B)} \quad (13)$$

To calculate the atom-pair similarity of two molecules, a set of atom-pair types is developed for each molecule. The atom-pair types are determined using the hydrogen-suppressed graph of the chemical structure and combining a simple atom typing scheme with the shortest path distances to arrive at the set of atom-pair types in the form, $\text{type}_i\text{-}d_{ij}\text{-}\text{type}_j$. The number of atom-pair types the two molecules share will determine their atom-pair similarities with 0.0 indicating no similarity and 1.0 indicating all atom pairs of the two molecules are shared. The atom-pair weights are all 1.0 by default though they may be fitted to bias important atom pairs. Weight fitting and application in Strike can only be done from the command line. See [Section 5.3.6 on page 81](#).

6.6.2 Similarity Measures in Descriptor Space

The four quantities defined here are measures of distance in descriptor space.

Manhattan Distance

$$\text{dist} = \sum_i w_i |x_i - x_i^{\text{probe}}| \quad (14)$$

The Manhattan distance metric, also known as the city-block distance, is a measure of the sum of geometric distances between points measured along axes at right angles. The distance being measured is summed over all variables. Put another way, the distance is calculated between all descriptors for a molecule and the probe value for each of those descriptors. For descriptor similarities, the probe value for each descriptor is the mean of the values of each probe molecule for that descriptor. Different weights for each descriptor, w_i , may be included only in command-line Strike calculations, otherwise all weights have the value of one. A value of zero indicates the probe molecule and test molecules are identical.

Euclidian Squared Distance

$$\text{dist} = \sum_i w_i (x_i - x_i^{\text{probe}})^2 \quad (15)$$

The Euclidean squared distance metric is a measure of the sum of geometric distances between points. The distance being measured is summed over all variables. Put another way, the distance is calculated between all descriptors for a molecule and the probe value for each of those descriptors. For descriptor similarities, the probe value for each descriptor is the mean of the values of each probe molecule for that descriptor. Different weights for each descriptor, w_i , may be included only in backend Strike calculations, otherwise all weights have the value of one. A value of zero indicates the probe molecule and test molecules are identical.

Euclidian Distance

The Euclidean distance is the square root of the expression in [Equation \(15\)](#).

Tanimoto Similarity

$$\text{dist} = \frac{\sum_i x_i x_i^{\text{probe}}}{\sum_i x_i x_i + \sum_i x_i^{\text{probe}} x_i^{\text{probe}} - \sum_i x_i x_i^{\text{probe}}} \quad (16)$$

The Tanimoto distance metric is a normalized measure of the similarity in descriptor space between a test molecule and a probe molecule. Similarities lie between one and zero with a value of one indicating identical molecules and a value of zero indicating completely dissimilar molecules.

Getting Help

Schrödinger software is distributed with documentation in PDF format. If the documentation is not installed in `$SCHRODINGER/docs` on a computer that you have access to, you should install it or ask your system administrator to install it.

For help installing and setting up licenses for Schrödinger software and installing documentation, see the *Installation Guide*. For information on running jobs, see the *Job Control Guide*.

Maestro has automatic, context-sensitive help (Auto-Help and Balloon Help, or tooltips), and an online help system. To get help, follow the steps below.

- Check the Auto-Help text box, which is located at the foot of the main window. If help is available for the task you are performing, it is automatically displayed there. Auto-Help contains a single line of information. For more detailed information, use the online help.
- If you want information about a GUI element, such as a button or option, there may be Balloon Help for the item. Pause the cursor over the element. If the Balloon Help does not appear, check that Show Balloon Help is selected in the Help menu of the main window. If there is Balloon Help for the element, it appears within a few seconds.
- For information about a panel or the folder that is displayed in a panel, click the Help button in the panel. The Help panel is opened and a relevant help topic is displayed.
- For other information in the online help, open the Help panel and locate the topic by searching or by category. You can open the Help panel by choosing Help from the Help menu on the main menu bar or by pressing CTRL+H.

To view a list of all available Strike-related help topics, click the Categories tab, then from the Categories menu, choose Strike. Double-click a topic title to view the topic.

If you do not find the information you need in the Maestro help system, check the following sources:

- *Maestro User Manual*, for detailed information on using Maestro
- *Maestro Command Reference Manual*, for information on Maestro commands
- *Ligand and Structure-Based Descriptors*, for information on descriptor generation
- Frequently Asked Questions pages at http://www.schrodinger.com/Strike_FAQ.html.

The manuals are also available in PDF format from the Schrödinger [Support Center](#). Information on additions and corrections to the manuals is available from this web page.

If you have questions that are not answered from any of the above sources, contact Schrödinger using the information below.

E-mail: help@schrodinger.com

USPS: Schrödinger, 101 SW Main Street, Suite 1300, Portland, OR 97204

Phone: (503) 299-1150

Fax: (503) 299-4532

WWW: <http://www.schrodinger.com>

FTP: <ftp://ftp.schrodinger.com>

Generally, e-mail correspondence is best because you can send machine output, if necessary. When sending e-mail messages, please include the following information, most of which can be obtained by entering `$SCHRODINGER/machid` at a command prompt:

- All relevant user input and machine output
- Strike purchaser (company, research institution, or individual)
- Primary Strike user
- Computer platform type
- Operating system with version number
- Strike version number
- Maestro version number
- mmshare version number

A

active ligands, extracting.....	55
activity property	35, 91
adding properties.....	33, 62
aqueous solubility	34
experimentally measured	36
aromatic proportion.....	33
atom-pair connectivity	51
atom-pair similarity.....	94
maximum.....	55
mean	55
atom-pair types.....	94
atom-pair weights.....	94
atoms, selecting.....	21–23
Auto-Help	28, 97

B

Balloon Help	28, 97
bivariate statistics	45, 46, 88
Build panel	19
Build QSAR Model panel.....	48
building structures.....	18–21
button menu	7

C

Calculate similarity panel	52
central tendency (of a variable).....	85
citing Strike in publications	1
city-block distance	95
command input area	31
Command Script Editor panel.....	24
command scripts— <i>see</i> scripts	
commands text box	31
conventions, document.....	vii
correlation coefficient	88
correlation matrix.....	88
covariance	88
cross validation	92
csv format.....	77
current working directory	4, 30

D

data reduction.....	48
database	
for calculating similarities	54

seeding.....	53
dependent descriptor	91
dependent variable	35, 91
descriptor space.....	51, 94
descriptors	32
from ligparse	32
from QikProp.....	32
directory	
current working	4, 25
local working	29
output.....	25
tutorial.....	30
distance	
Euclidean	95
in descriptor space	94
Manhattan	95
Tanimoto.....	95

E

eigenvalue	48
entries, Project Table.....	11
including, excluding, and fixing	16
selecting	15
sorting	13, 55
environment variables	
DISPLAY	4
SCHRODINGER	3–4
ePlayer.....	13, 14
Euclidean distance	95
Euclidean similarity	58
Euclidean squared distance	95
Euclidean squared similarity.....	58
excluded entries	16
extracting actives.....	55

F

failed job	41
file I/O directory.....	25, 30
filters, project entry	15
Fisher skewness.....	87
fixed entries	16
format, csv.....	77
fragments, building structures from	18
F-statistic.....	93
full screen mode.....	6, 11
function key macros— <i>see</i> scripts	

G

grow bond 19

H

Help panel 28, 97

I

included entries 16

independent descriptors 91

independent variables..... 35, 91

input file 77

installation..... 29

inverting selection 41, 65

J

jobs, running in Maestro 26–27

K

Kendall tau 90

kurtosis 88

L

leave-group-out (LGO) validation 92

leave-n-out validation..... 92

lgoCycles..... 93

lgoPercent 93

ligparse utility..... 32

Local Correlation Integral Outlier (LOCI)

Detection..... 93

log file, saving Maestro..... 28

Mmacros—*see* scripts

Maestro

main window 4, 5

menus..... 6

quitting..... 28

running jobs from 26–27

scratch projects 11

starting 4

undoing operations 26

main window..... 5

Manhattan distance 95

Manhattan similarity 58

Max AP Similarity 55

maximum atom-pair similarity 55, 58

mean (of variable x) 85

mean absolute deviation..... 86

Mean AP Similarity 55

mean atom-pair similarity 58

mean squared deviation..... 87

measured aqueous solubilities..... 36

median 85

menu button 7

MLR optimal subset (MLRO) method..... 50

mode..... 86

model-building, PLS 34

molecular properties..... 32

from QikProp..... 32

Monitor panel..... 27

mouse functions 3

Project Table panel 16–17

Workspace 10

multiple linear regression (MLR) 49, 92

N

New Plot dialog box..... 43, 66

n-factor predictor 38, 48

non-carbon proportion 33

normal distribution 87, 88

number of data points..... 85

O

online help..... 28

optimal set of descriptors 49

optimal subset 49

outlier detection 93

output file, model-building job 40

P

partial least squares (PLS) 91

Partial Least Squares method..... 34

Pearson r..... 89

plot name..... 43, 66

plot series name..... 43, 66

plot settings 45

Plot XY panel..... 38, 43, 66

Plot XY panel features 40

plot, selected 44

-
- Predict based on QSAR Model panel 42
 - Predicted Activity property 36, 42
 - predicted values 42
 - prediction, running 42
 - predictor 37
 - Preferences panel 25, 26, 30
 - principal component analysis (PCA) 48, 91
 - probe molecules 54
 - product installation..... 97
 - project entries, *see* entries, Project Table
 - Project Facility, introduction..... 11
 - Project Table 32
 - adding or removing properties from..... 47
 - adding properties 33, 62
 - Project Table panel..... 13
 - menus..... 14
 - mouse functions..... 16–17
 - shortcut keys 17
 - projects 11
 - properties..... 32
 - adding 33, 62
 - from ligparse..... 32
 - in Statistics panel..... 47
 - p-value..... 93
 - Python scripts—*see* scripts
- Q**
- QikProp..... 32
 - QSAR model
 - building..... 34, 62
 - viewing results 40
 - q-squared..... 93
 - quitting Maestro 28
- R**
- random selection 33
 - regression methods..... 90
 - response variable..... 91
 - results table 37
 - root mean squared deviation 87
 - rotatable bonds 36
 - r-squared 89
- S**
- scatter plot..... 89
 - Schrödinger contact information..... 98
 - scratch entries..... 12
 - scratch projects..... 11
 - scripts
 - function key macros..... 25
 - macros..... 25
 - Maestro command 24
 - Python..... 23
 - seeding database..... 53
 - select by properties 41, 65
 - selected plot 44
 - selecting objects in the Workspace 7, 21
 - selection, inverting 41, 65
 - shortcut keys
 - main window 11
 - Project Table panel 17
 - similarity 51
 - 2D structure space 51
 - atom-pair..... 51
 - descriptor space 51
 - maximum atom-pair 58
 - mean atom-pair 55, 58
 - skewness..... 87
 - Snedecor and Cochran kurtosis..... 88
 - Sort Project Table panel 56
 - sorting of entries 55
 - Spearman rho 89
 - standard deviation 86
 - statistics
 - bivariate 45, 88
 - univariate 45, 85
 - statistics script..... 45
 - strike command..... 77
 - Strike input file..... 77
 - Strike Univariate and Bivariate Statistics panel
 - closing and reopening..... 47
 - Strike, citing use of 1
 - structure file
 - active thermolysin ligands 52
 - decoy ligands 52
 - structures
 - building 18–21
 - displaying in sequence..... 13
- T**
- Tanimoto distance 95
 - Tanimoto similarity 58
 - technical support 28

test set..... 33
 creating 41, 64
toolbar
 Build panel..... 20–21
 main window 7–10
 Project Table panel 13–14
total variance..... 48
training set..... 33
 selected 41, 65
tutorial directories 30
two-dimensional structure space..... 51

U

undoing Maestro operations..... 26
univariate statistics 45, 46, 85

V

validation..... 41, 64
variance 86
 total..... 48
View QSAR Model dialog box 40

W

working directory 29
Workspace
 description 4
 full screen mode 6, 11
 including, excluding, and fixing entries 16
 mouse functions..... 10
 scratch entries 12

120 West 45th Street
32nd Floor
New York, NY 10036

101 SW Main Street
Suite 1300
Portland, OR 97204

3655 Nobel Drive
Suite 430
San Diego, CA 92122

Dynamostraße 13
68165 Mannheim
Germany

QuatroHouse, Frimley Road
Camberley GU16 7ER
United Kingdom

SCHRÖDINGER.